



Evaluación de Modelos de Machine Learning para la Predicción de Crímenes en la Ciudad de Medellín

Victor Daniel Muñoz Jaramillo

Universidad Nacional de Colombia
Facultad de Minas, Departamento de Ciencias de la Computación y de la Decisión
Medellín, Colombia
2021



Machine Learning Models for Crime Prediction in Medellin City

Victor Daniel Muñoz Jaramillo

National University of Colombia
Faculty of Mines, Department of Computer Science
Medellin, Colombia
2021

Evaluación de Modelos de Machine Learning para la Predicción de Crímenes en la Ciudad de Medellín

Victor Daniel Muñoz Jaramillo

Tesis presentada como requisito parcial para optar al título de:
Magister en Ingeniería – Analítica

Directora:

Mónica Ayde Vallejo Velásquez Ph.D

Codirector:

José Édinson Aedo Cobo Ph.D

Universidad Nacional de Colombia

Facultad de Minas, Departamento de Ciencias de la Computación y de la Decisión
Medellín, Colombia

2021

Agradecimientos

En primer lugar agradezco muy especialmente a la Doctora Mónica Ayde Vallejo Velásquez por su excelente orientación, enseñanzas y acompañamiento durante todo el desarrollo de la tesis. Igualmente al Dr. José Edinson Aedo, co-asesor, por su guía durante todas las etapas del trabajo.

También agradezco al proyecto de Minciencias '*Administración Inteligente de Problemas de Seguridad Ciudadana a través de Modelos y Herramientas generadas a partir de Plataformas para Territorios Inteligentes apoyadas por Estrategias de Participación Ciudadana en la Ciudad de Medellín*' (BPIN-2020000100044), en el marco del cual se definió el tema de investigación.

Finalmente quiero manifestar mis agradecimientos a mi familia y a todas las personas que de alguna manera me apoyaron y animaron en la realización de este trabajo.

Resumen

La seguridad ciudadana se ha convertido en una de las principales preocupaciones de los gobiernos dada su relación directa con la calidad de vida de las personas, el crecimiento económico y el desarrollo de las regiones. Por su parte, el crimen se ha constituido como uno de los principales factores que afecta la seguridad, y para combatirlo, los gobiernos han asignado una cantidad de recursos que se podrían utilizar para proyectos de inversión como los de infraestructura. Históricamente el enfoque de las estrategias de las autoridades locales se ha centrado en respuestas reactivas como la captura de los delincuentes, no obstante, recientemente se ha reconocido la necesidad de desarrollar estrategias preventivas de vigilancia y control de los espacios públicos, mediante el uso de tecnologías de aprendizaje automático (*Machine Learning*). Por esta razón, con el fin de colaborar con las estrategias de las autoridades para la gestión de los recursos, en esta tesis de maestría se realiza la evaluación de tres modelos de *Machine Learning* para la predicción del crimen en la ciudad de Medellín: un clasificador de bosques aleatorios, un modelo de regresión logística y una máquina de vectores de soporte (SVM, de sus siglas en inglés *support vector machine*). La metodología implementada integra el resultado de estudios anteriores con el proceso estándar de la industria para la minería de datos (CRISP-DM, de sus siglas en inglés *Cross Industry Standard Process for Data Mining*) como una estrategia general de resolución de problemas de la unidad de estudio. Como parte de la metodología, inicialmente se realiza un entendimiento y descripción de diferentes fuentes de información disponibles en la ciudad de Medellín. Luego, a partir de la identificación de los datos, su preparación y análisis, se formulan los modelos para la predicción de zonas calientes con información histórica del número de incidentes e información de la tasa de desempleo en la ciudad. Específicamente los modelos se construyen para la predicción del hurto a personas en las modalidades de atraco, descuido, cosquilleo y raponazo. Finalmente, el desempeño de los tres modelos se compara contra un modelo basado en reglas, y se evalúan en términos de la exactitud, exhaustividad/sensibilidad (*recall*), precisión y el valor F1.

Palabras clave: Predicción del crimen, zonas calientes, *machine learning*, modelos predictivos, predicción de crimen en Medellín, seguridad ciudadana.

Abstract

Public safety is one of the main concerns of governments, given its direct relationship with people's wellbeing, economic growth, and the development of the regions. For its part, crime has been detected as one of the main factors that affect the feeling of security, assigning it a considerable percentage of government resources to combat it. Historically, national authorities' strategies have focused on reactive responses such as the capture of criminals, however, the need to develop preventive strategies for surveillance and control of public spaces has been recently recognized. For this reason, in order to improve the strategies currently used by the authorities for resource management, this master's thesis evaluates three Machine Learning Models: a random forest classifier, a logistic regression model, and a support vector machine (SVM), for the prediction of crime in the city of Medellin. The proposed methodology integrates previous studies that have been conducted in other regions with the Cross Industry Standard Process for Data Mining (CRISP-DM) as a general strategy for problem solving of the unit of study. As part of the methodology, it begins with the understanding and description of the available information in the city of Medellin. Then, from the identification of the data, its preparation, and analysis, the Machine Learning models are formulated for the prediction of crime hotspots, using the information about historical incidents and the unemployment rate. Finally, the performance of the 3 models is evaluated in terms of accuracy, recall, precision, and F1 score, and each of the models is compared with the result obtained by using a base model built on rules that the authorities could establish.

Keywords: Crime prediction, hotspot prediction, machine learning, predictive models, crime prediction in Medellin, public safety.

Contenido

Agradecimientos	v
Resumen	vii
1 Introducción	2
1.1 Motivación	4
1.2 Alcance	5
1.3 Objetivos	6
1.4 Descripción General de la Tesis	6
2 Marco teórico	8
2.1 Caracterización del Crimen	8
2.2 Machine Learning para la Predicción del Crimen	10
2.2.1 Categorías de Machine Learning	11
2.2.2 Clasificación de Algoritmos de tipo Machine Learning	11
2.2.3 Métricas de Rendimiento	13
2.2.4 Métodos de validación	16
2.3 CRISP-DM	17
3 Antecedentes y Estado del Arte	20
4 Metodología para la Predicción de Crímenes	26
4.1 Desarrollo de la Metodología	27
5 Análisis del Problema y Recolección de Datos	32
5.1 Comprensión del tema de interés	32
5.2 Estudio y Comprensión de los Datos	33
5.2.1 Fuentes de Información	33
5.2.2 Descripción y Exploración de los Datos	36
5.3 Preparación de los Datos	47
5.3.1 Definiciones	47
5.3.2 Selección de los datos	48
5.3.3 Preprocesamiento de los datos	48
5.3.4 Construcción e integración de información	49

6 Modelos para la Predicción del Crimen en la Ciudad de Medellín	55
6.1 Selección de la Técnica de Modelado	55
6.2 Técnica de Evaluación	55
6.3 Construcción del Modelo	56
6.4 Resultados	57
7 Conclusiones y Recomendaciones	62
7.1 Conclusiones	62
7.2 Recomendaciones	62
Bibliografía	64

Lista de Figuras

2-1	Número de delitos por cada 100 mil habitantes para el año 2020 en Colombia. Elaborado a partir de [1]	9
2-2	Matriz de Confusión	13
2-3	Fases de CRISP-DM. Elaborado a partir de [2]	18
2-4	Fases de CRISP-DM detallado. Elaborado a partir de [2]	19
5-1	Recopilación de los registros de hurto a persona.	34
5-2	Recopilación de los polígonos de las comunas de Medellín.	34
5-3	Ingreso al portal SIATA.	35
5-4	Consulta de variables meteorológicas SIATA.	35
5-5	Recopilación de indicadores demográficos DANE.	36
5-6	Comportamiento anual y mensual de los incidentes en la ciudad de Medellín.	37
5-7	Comportamiento semanal de los incidentes en la ciudad de Medellín.	38
5-8	Comportamiento por hora de los incidentes en la ciudad de Medellín durante los días laborales.	38
5-9	Comportamiento por hora de los incidentes en la ciudad de Medellín durante el fin de semana.	38
5-10	Información demográfica de las víctimas.	39
5-11	Información de las características del hurto en Medellín entre 2003 y 2020. .	40
5-12	Información del bien robado.	42
5-13	Distribución espacial de los crímenes (a) por comuna y (b) por barrio.	42
5-14	Tasa de desempleo en Medellín entre el enero 2000 y diciembre de 2020. . . .	43
5-15	Comparación de la tasa de desempleo para en Medellín.	43
5-16	Humedad y precipitación en la ciudad de Medellín	45
5-17	Presión Atmosférica y Velocidad del Viento en la ciudad de Medellín	46
5-18	Temperatura en la ciudad de Medellín	46
5-19	Escala espacial disponible para la ciudad de Medellín	49
5-20	Distribución de hurto en la ciudad de Medellín entre el 2015 y 2019	51
5-21	Distribución de hurto en la ciudad de Medellín entre el 2015 y 2019	51
5-22	Comparación para diferentes valores del umbral y área cubierta de la ciudad.	53
5-23	Distribución de zonas calientes por celda	54

Lista de Tablas

1-1	Información de Criminalidad - Número de Delitos por 100 mil habitantes [1]	3
1-2	Indicadores objetivos sobre la seguridad en Medellín [3, 4]	4
3-1	Compración de los estudios realizado en torno a la predicción del crimen (ND = No Definido) [5, 6, 7, 8, 9, 10, 11, 12, 13, 14].	22
4-1	Parámetros que se deben reportar en la predicción del crimen [15].	26
4-2	Metodología para la predicción del crimen con modelos de Machine Learning	27
5-1	Campos que componen la base de datos de hurtos	37
5-2	Clasificación de los bienes robados	41
5-3	Calidad de los datos meteorológicos	44
5-4	Estadísticas descriptivas de las variables meteorológicas	45
5-5	Estadísticas descriptivas de la distribución de celdas	51
5-6	Comparación de diferentes valores para la ventana temporal	52
6-1	Resultados de la Validación Cruzada Doble sobre Regresión Logística	57
6-2	Resultados de la validación cruzada doble sobre bosques aleatorios	58
6-3	Resultados de la validación cruzada doble sobre SVC	59
6-4	Métricas de los modelos de predicción	59
6-5	Resumen de la metodología aplicada	61

1 Introducción

La seguridad ciudadana ha sido reconocida a nivel mundial como uno de los grandes pilares para la prosperidad de las naciones, ya que la promoción y el mantenimiento de entornos seguros son actividades indispensables para el desarrollo económico y social de una región. Como lo indica la pirámide de Maslow en la teoría de la Motivación, los deseos más elevados de las personas solo son alcanzados cuando se satisfacen las necesidades primarias como la seguridad física [16].

Asimismo, organizaciones internacionales como la Organización para la Cooperación y el Desarrollo Económicos (OCDE) consideran la seguridad individual como un factor determinante para el bienestar de las personas, por lo cual constituye un reto para las administraciones públicas mejorar los niveles de seguridad de los ciudadanos [17]. Por otro lado, el Programa de las Naciones Unidas para el Desarrollo, define la seguridad ciudadana como “el proceso de establecer, fortalecer y proteger el orden civil democrático, eliminando las amenazas de violencia en la población y permitiendo una coexistencia segura y pacífica”, indicando que es un factor importante que se debe tener en cuenta para mejorar la calidad de vida de la población [18].

A nivel nacional, a través del Ministerio de Defensa de Colombia [19], se define la seguridad ciudadana como “el conjunto de acciones integrales que busca proteger de manera efectiva a las personas, de los delitos y de los comportamientos que afectan su integridad”, y por medio de varias encuestas, se ha resaltado que los principales factores que atentan contra la seguridad de los colombianos se atribuyen a prácticas sociales en espacios públicos (tales como el consumo de bebidas alcohólicas y sustancias psicoactivas), porte de armas blancas, déficit en las habilidades sociales (ej. manejo de las emociones, inadecuadas pautas de crianza), déficit en la cultura de la legalidad por parte de los ciudadanos, y la presencia de Grupos de Delincuencia Común Organizada [19, 20]. Adicionalmente, pese a que en los últimos años se han conseguido importantes logros en la lucha contra los delitos, se ha remarcado la necesidad por mejorar las políticas en torno a la seguridad, para disminuir las actividades ilícitas que constituyen un problema para los ciudadanos y para el crecimiento económico.

En el ámbito jurídico, actualmente el Código Penal Colombiano contempla un total de 363 delitos en la modalidad de tipos básicos, sin contar aquellos que son agravantes o atenuantes [21]. Sin embargo, dentro de las principales actividades delictivas que afectan la seguridad

ciudadana y aquellas que poseen las mayores tasas son: el homicidio común, las lesiones personales y las diferentes modalidades de hurto y extorsión. Como se puede observar en la Tabla 1-1, aunque estas modalidades tuvieron una disminución considerable para el año 2020, no se debe olvidar que por la emergencia sanitaria por COVID-19, los gobiernos impusieron jornadas de cuarentena, reduciendo el número de ciudadanos expuestos a los peligros en la calle. En contraste, si se comparan los años anteriores, se puede observar que delitos como el homicidio, el hurto y la extorsión se encontraban en aumento.

Tabla 1-1: Información de Criminalidad - Número de Delitos por 100 mil habitantes [1]

Delitos	2017	2018	2019	2020
Homicidio total	24,9	25,9	25,7	24,3
Lesiones personales	269,1	277,5	236,8	167,5
Hurto a personas	425,6	515,8	609,1	408,9
Extorsión	11,2	14,1	16,6	16,1
Secuestro	0,4	0,4	0,2	0,3

Por otro lado, a nivel departamental y específicamente en la ciudad de Medellín se ha demostrado que los delitos contra el patrimonio económico, especialmente el atraco en vía pública, constituyen el mayor porcentaje de casos que aporta al nivel de victimización [3]; y aunque las cifras para el 2020 indican una disminución del número de homicidios, hurtos a personas y muertes violentas [4], este comportamiento coincide con el periodo de confinamiento restrictivo por las medidas tomadas por el gobierno para enfrentar la pandemia por COVID-19. Solo para la violencia intrafamiliar, estas restricciones se asocian a un aumento de denuncias para el 2020. Sin embargo, realizando un análisis sobre el comportamiento del crimen antes de la pandemia, se puede observar en la Tabla 1-2 que para el 2018 se tuvo un incremento del 3% en la tasa de homicidios y un una variación del 16% en la denuncia de hurtos a personas, mientras que para el año 2019 se reportó un aumento del 21% en la denuncia de hurtos a personas. Finalmente, encuestas realizadas mostraron que Medellín solo contaba con un 41% de las personas sintiéndose seguras en la ciudad para el 2018 [3], y a pesar del valor de la métrica, no se ha evidenciado una iniciativa por parte de las autoridades para mejorar este indicador, ya que para diciembre de 2020 se reportó un 33% [4].

A nivel latinoamericano se vienen implementando diversas estrategias para tratar de gestionar y mejorar los índices de seguridad ciudadana, entre ellas, la integración de la participación de la comunidad y coordinación con las autoridades [22]. La seguridad y su gestión, vistos como uno de los factores importantes para el bienestar, han permitido la incorporación de nuevas perspectivas desde la introducción del concepto de ciudades inteligentes en sus diversas visiones. Es importante anotar, que si bien el término ciudad inteligente es utilizado extensamente, todavía no existe una definición coherente que tenga una amplia aceptación [23, 24, 25]. Existen diferentes perspectivas desde diversas disciplinas tales como sociología,

Tabla 1-2: Indicadores objetivos sobre la seguridad en Medellín [3, 4]

Indicador	2017	2018	2019	2020
Tasa de homicidios por cien mil habitantes	23.2 %	26.1 %	23.8 %	14.4 %
Número de víctimas de violencia intrafamiliar	4856	5000	5506	9091
Número de denuncias por hurto a personas	17709	21079	26700	17636
Tasa de muertes violentas por cien mil habitantes	47.5 %	49.0 %	47.7 %	33.8 %
Número de denuncias de hurtos de carros y motos	5010	5705	5784	5092

informática (TIC), estudios urbanos, salud pública, entre otros. Desde una visión tecnológica, el adjetivo “inteligente” (*Smart*), hace referencia en la mayoría de los casos, al uso de la TIC, para mejorar la eficiencia de los servicios desplegados en una ciudad. En este contexto, la seguridad y protección de los ciudadanos se considera un servicio básico en los modelos de ciudad inteligente [26]. La disponibilidad de información real (bases de datos suministradas por las autoridades) sobre comportamiento del crimen y sobre la percepción de seguridad por parte de la ciudadanía, así como nuevas herramientas para apoyar procesos de análisis, predicción y toma de decisiones como la inteligencia artificial y *big data*, están abriendo nuevas perspectivas en el manejo de la seguridad ciudadana.

Asimismo, en la literatura se puede encontrar el desarrollo de modelos de predicción de delitos en la ciudad, donde destacan los modelos de aprendizaje automático, los modelos autorregresivos y los algoritmos de aprendizaje profundo. Y dependiendo del tipo de tarea (ej: clasificación, regresión, etc.), se pueden observar diferentes enfoques, tales como: la predicción de zonas calientes, la predicción del número y/o tasa de delitos a lo largo del tiempo, o la predicción de la categoría del delito [15]. Estos modelos buscan un objetivo común, que consiste en mejorar las estrategias para la gestión de los recursos de las autoridades.

1.1. Motivación

Históricamente el enfoque de las estrategias de las autoridades nacionales se ha centrado en la captura de los delincuentes, no obstante, recientemente se ha reconocido que no basta solo con la observación y reacción por parte de las autoridades, sino que se requiere desarrollar una estrategia preventiva de vigilancia y control de los espacios públicos, municipios y veredas [19]. Para lograrlo, la Policía Nacional de Colombia ha ejecutado ideas innovadoras como el desarrollo de capacidades de inteligencia artificial y la promoción de la participación cívica, con el fin de anticipar la comisión de los delitos y proteger a los ciudadanos. En línea con esto, se ha dotado con diferentes mecanismos para agilizar los tiempos de respuesta ante incidentes, como: cámaras de reconocimiento facial, cámaras focalizadas en puntos críticos de las ciudades, cámaras para la identificación de vehículos y reconocimiento de placas, y drones para la vigilancia de distritos, municipios y veredas [19].

Sin embargo, se observa que a pesar de los esfuerzos por parte de la ciudad de Medellín y de la nación en el fortalecimiento de los Sistemas de Información de Seguridad y Convivencia [27], la cantidad de dispositivos adquiridos por las autoridades para atención oportuna de incidentes [28, 29] y el desarrollo de canales para promover la participación ciudadana en ámbitos de la seguridad [30], no se percibe una mejora considerable en los indicadores de los delitos contra la vida y contra el patrimonio económico, la percepción de la seguridad en la ciudad, ni en la cultura de denuncia por parte de las víctimas [3].

Finalmente, dentro de las motivaciones para el desarrollo de esta tesis se encuentra la disponibilidad de diversas fuentes de información, las cuales pueden articularse en la construcción de modelos de *Machine Learning* para la predicción de delitos en la ciudad. La alcaldía de Medellín, por su parte, cuenta con una base de datos sobre delitos y accidentes, y sobre los mapas espaciales de las unidades administrativas de la ciudad; el SIATA (Sistema de Alerta Temprana del Valle de Aburrá) almacena información sobre variables meteorológicas como temperatura, humedad, velocidad del viento y precipitaciones; y el DANE, el departamento de estadísticas de la nación, publica periódicamente cifras demográficas como el crecimiento de la población y la tasa de desempleo.

1.2. Alcance

La presente tesis de maestría propone la articulación de diversas fuentes de información, y la evaluación de varios modelos de *Machine Learning* para la predicción del crimen en la ciudad de Medellín. Específicamente se espera integrar información sobre los hurtos a personas en sus modalidad de atraco, descuido, cosquilleo y raponazo, con variables meteorológicas e indicadores demográficos. Y como resultado, se espera obtener información relevante para ayudar a las autoridades a formular estrategias para la toma de decisiones y para tener una respuesta rápida ante los incidentes de seguridad.

Se pretende integrar los estudios recopilados en torno a la predicción del crimen con la metodología denominada Proceso Estándar de la Industria para la Minería de datos (CRISP-DM, por sus siglas en inglés), y desarrollar cada uno de los puntos que la integran, pasando por la descripción y evaluación de la situación, el entendimiento de las fuentes de información, la preparación de los datos, la construcción de los modelos y concluyendo con la evaluación de los resultados obtenidos. Los resultados de esta tesis se utilizarán como insumo para el despliegue de una solución en forma de reporte o aplicación web en el desarrollo del proyecto macro de Minciencias BPIN-2020000100044.

1.3. Objetivos

Objetivo general

Evaluar modelos de *Machine Learning* para la predicción de crímenes en la ciudad de Medellín a través de datos de fuentes de información públicas y privadas de la región.

Objetivos específicos

- Identificar las fuentes de información públicas y privadas que puedan contribuir a la predicción del crimen en la ciudad de Medellín.
- Identificar, integrar y validar la calidad de las fuentes de información que contribuyan a la predicción del crimen.
- Priorizar los tipos de crímenes a analizar considerando las bases de datos existentes.
- Implementar varios modelos de *Machine Learning* orientados a la predicción de patrones de crímenes.
- Evaluar y comparar el desempeño de los modelos implementados.

1.4. Descripción General de la Tesis

La tesis se encuentra organizada en siete capítulos guiados al cumplimiento de los objetivos planteados, y se estructuran como sigue:

En el capítulo 2 se presenta una contextualización sobre el crimen, específicamente sobre la concepción que se tiene del hurto y robo, tanto a nivel internacional como nacional. Se da un entendimiento sobre los conceptos de aprendizaje automático, en donde se incluyen los modelos, las métricas para medir el desempeño y los métodos de evaluación. Finalmente, se presenta una introducción a la metodología CRISP-DM, con una breve descripción de las principales actividades.

El capítulo 3 realiza una descripción de los estudios encontrados en la literatura, exponiendo los métodos utilizados para la predicción del crimen, las métricas para la medición del desempeño de los modelos y las estrategias de validación más utilizadas.

El capítulo 4 propone una metodología para la predicción del crimen en la ciudad de Medellín, que integra los puntos de la metodología CRISP-DM y parámetros específicos sobre la predicción del crimen que se encuentra dentro de la literatura. Esta se define como una metodología para la reproducción de los resultados en el presente trabajo y como un modelo

para que futuros estudios la puedan integrar en sus investigaciones.

En el capítulo 5 se presenta el desarrollo del problema de investigación en torno a la metodología propuesta, pasando por el planteamiento de los objetivos en términos del aprendizaje automático, la enumeración de las fuentes de información disponibles, así como la descripción y exploración de los datos y finalmente, la preparación de los datos, en donde se detalla la selección de las características y las fases de preprocesamiento, construcción e integración.

Por su parte, el capítulo 6 contiene el desarrollo de los modelos, abordando la estrategia de validación, las métricas para la medición del desempeño, la calibración de los hiperparámetros seleccionados y el análisis de los resultados.

Finalmente en el capítulo 7 se exponen las conclusiones y recomendaciones para trabajos futuros.

A partir de este trabajo se ha logrado:

Dos artículos publicados en conferencia:

- V. Munoz, M. Vallejo and J. E. Aedo, Exploratory Analysis of Crime Behavior in the City of Medellin, 2021 2nd Sustainable Cities Latin America Conference (SCLA), 2021, pp. 1-5, doi: 10.1109/SCLA53004.2021.9540095.
- V. Muñoz, M. Vallejo and J. E. Aedo, Machine Learning Models for Predicting Crime Hotspots in Medellin City, 2021 2nd Sustainable Cities Latin America Conference (SCLA), 2021, pp. 1-6, doi: 10.1109/SCLA53004.2021.9540132.

2 Marco teórico

Con el objetivo de brindar un entendimiento sobre los conceptos, herramientas y metodologías que se utilizan a lo largo de la presente tesis, este capítulo introduce las definiciones en torno al crimen en la ciudad, el aprendizaje automático y la metodología CRISP-DM.

2.1. Caracterización del Crimen

Según la Real Academia Española se define el delito como una acción u omisión voluntaria o imprudente penada por la ley [31], mientras que la palabra crimen suele utilizarse para aquellos casos en los que el acto ilegal es más grave [32]. En el idioma español, estas dos palabras suelen utilizarse como sinónimos, pero en realidad, semánticamente todo crimen es un delito, pero no todo delito es un crimen. Por su parte, en el idioma inglés suele utilizarse el mismo término (*crime*) para referirse a estas dos palabras, y dado que gran parte de la literatura utilizada se encuentra en este idioma, se decide utilizar indistintamente estos conceptos para la presente tesis.

Como punto de partida para la caracterización del crimen, se realiza una revisión del Código Penal Colombiano [21], un documento en donde se tipifican los delitos y se determinan las penas que le corresponden. Actualmente, el código contempla un total de 363 delitos agrupados en 19 categorías, entre las cuales se pueden destacar: delitos contra la vida y la integridad personal, delitos contra las personas y bienes, delitos contra la libertad individual y delitos contra la familia. Si bien, la predicción de delitos podría plantearse en cada una de estas categorías, esta investigación se enfoca en modelos de predicción tomando para el análisis el hurto a personas (en sus modalidad de atraco, descuido, cosquilleo y raponazo), ya que este delito se posiciona como la categoría con mayores índices a nivel nacional [3]. Esta proporción se puede visualizar en la Figura 2-1.

Aunque coloquialmente los conceptos de hurto y robo se utilizan indistintamente para referirse a la misma acción, se puede encontrar que la Real Academia Española hace una distinción de los conceptos indicando que el robo, a diferencia del hurto, es un delito en donde se emplea la violencia o intimidación sobre las personas, o fuerza en las cosas [33, 34].

Adicionalmente, el estándar internacional para el registro de incidentes NIBRS (Sistema Nacional de Información Basado en Incidentes), define ambos conceptos como sigue [35]:

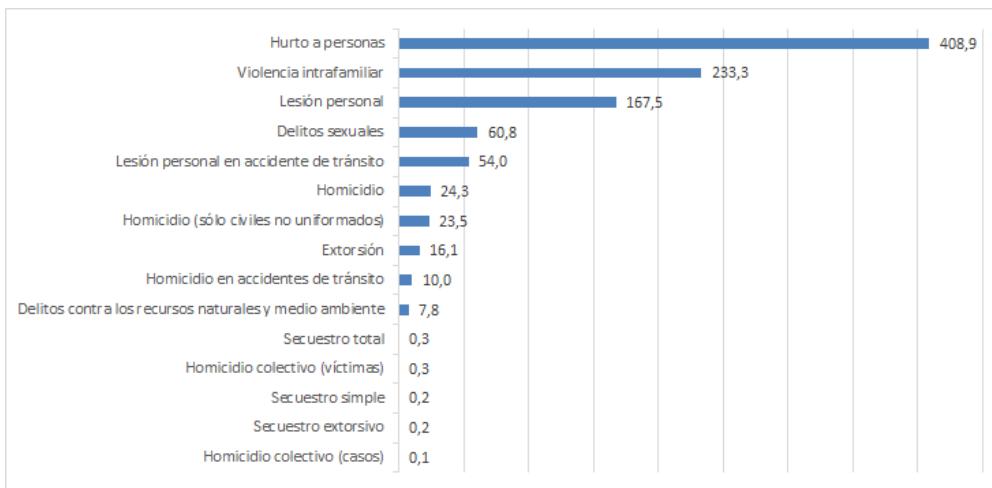


Figura 2-1: Número de delitos por cada 100 mil habitantes para el año 2020 en Colombia.
Elaborado a partir de [1].

1. **Hurto:** adquirir, transportar, llevarse bienes que le pertenecen legalmente a otra persona. Dentro de este delito se encuentra:
 - **Ratería:** la apropiación ilegal de artículos que se encuentran en posesión física de otra persona, en donde la víctima normalmente no se entera inmediatamente del hurto.
 - **Hurto de Carteras:** el agarrar o arrebatar una cartera, un bolso, etc. de la posesión física de otra persona. Para este caso si se utiliza más fuerza de lo que es realmente necesario para arrebatar la cartera de la mano, o si la víctima se resiste de cualquier manera, entonces ha ocurrido un robo.
2. **Robo:** llevarse o intentar llevarse cualquier objeto de valor, en circunstancias de confrontación, que se encuentra bajo el control, custodia o cuidado de otra persona mediante la fuerza o amenaza o violencia y/o poner a la víctima en temor de daño inmediato.

Por otro lado, el Código Penal Colombiano abarca ambos conceptos dentro de los artículos 239 y 240 [21], definiendo el hurto como la posesión de una cosa mueble ajena, con el propósito de obtener provecho para sí o para otro, y señala que si se realiza con violencia o colocando a la víctima en condiciones de indefensión o inferioridad, se debe establecer como Hurto Calificado. Adicionalmente, en la encuesta de Convivencia y Seguridad Ciudadana, el Departamento Administrativo Nacional de Estadística (DANE) incluye las siguientes definiciones dentro del hurto a personas [36]:

- **Cosquilleo:** modalidad en donde los delincuentes aprovechan los tumultos en centros comerciales, transporte masivo o en otros lugares públicos, para extraer dinero, celulares u otros elementos, sin que la víctima se percate de lo ocurrido. Los objetos

son sustraídos de morrales, maletas, bolsos, chaquetas, bolsillos, etc., sin que la otra persona se percate de ello.

- **Engaño:** Aquel hurto en el cual se utilizan la mentira, la falsedad y la suplantación para que las víctimas accedan de forma voluntaria a entregar sus objetos de valor.
- **Raponazo:** es el hurto mediante una acción rápida, por medio del cual se arrebata sus pertenencias a la víctima antes de que esta pueda reaccionar.
- **Atraco:** es el robo con intimidación empleando armas blancas, armas de fuego o contundentes para obligar a la víctima a entregar sus pertenencias.
- **Fleteo:** modalidad de hurto que se realiza contra las personas que salen de las entidades financieras y acaban de realizar retiros en efectivo. A la salida del banco, son objeto de seguimiento por varios individuos que los abordan, amenazan y arrebatan el dinero.
- **Paseo Millonario:** forma de hurto en la cual la víctima aborda un vehículo donde es retenida por los delincuentes que luego la fuerzan a acompañarlos a distintos cajeros automáticos y entregarles dinero. Por ejemplo, un taxista recoge a una persona en la calle, más adelante otros delincuentes ingresan al taxi, lo amenazan y lo llevan a un cajero automático para robarle su dinero.
- **Descuido:** el delincuente aprovecha la falta de atención de la víctima a la hora de cuidar sus pertenencias para arrebatárselas sin que esta se dé cuenta.

2.2. Machine Learning para la Predicción del Crimen

En la búsqueda por mejorar las estrategias de las autoridades para combatir cada uno de los crímenes descritos, el aprendizaje automático se ha posicionado como una de las herramientas más utilizadas para el descubrimiento de patrones en el comportamiento de los actos delictivos. Este mecanismo, utiliza los datos a través de algoritmos computacionales, con el fin de tomar acción como la clasificación de la información en diferentes categorías [37]. Entre los diferentes acercamientos para el reconocimiento de patrones destaca el Aprendizaje Automático o *Machine Learning*. *Machine Learning* se puede definir como la rama de estudio que le brinda a los computadores la habilidad de aprender sin contar con una programación explícita [38]. Se utiliza principalmente para simplificar problemas que requieren numerosos ajustes manuales o largas listas de reglas, y para obtener información sobre problemas complejos y con grandes cantidades de datos [39, p. 7].

Al ejecutar un algoritmo de tipo *Machine Learning*, se obtiene como resultado un modelo $y(x)$, que toma las entradas x de un conjunto de datos y que genera una salida y . La forma de $y(x)$ se determina durante la denominada fase de entrenamiento o aprendizaje, utilizando

los datos de entrenamiento. Y una vez el modelo está entrenado, se realiza una fase de prueba o validación en donde se le entrega al modelo un conjunto nuevo de datos para determinar su salida [37].

2.2.1. Categorías de Machine Learning

Dependiendo de la cantidad de supervisión humana que se le debe suministrar al algoritmo, las técnicas de *Machine Learning* se pueden dividir en:

Aprendizaje Supervisado

Los algoritmos de aprendizaje supervisado utilizan conjuntos de datos previamente etiquetados para entrenar los modelos y encontrar la solución deseada [39, p. 8][40]. Esta categoría comprende varias tareas, entre las cuales se pueden destacar [37]:

- Tarea de Clasificación: los modelos que hacen parte de esta familia deben producir una función capaz de asignar una o más categorías a un conjunto de entrada.
- Tarea de Regresión: conceptualmente es similar a la tarea clasificación, pero en su lugar, la salida se encuentra en un dominio continuo y no discreto.

Aprendizaje No Supervisado

Tiene lugar cuando no se dispone de datos etiquetados para el entrenamiento. Para estos algoritmos, lo único que se conoce son los datos de entrada y el objetivo consiste en encontrar un tipo de organización de los datos que permita simplificar su análisis. Los algoritmos más habituales son: Algoritmos de Clustering, Análisis de Componentes Principales y la Deteción de Anomalías [39, p. 10].

Aprendizaje por refuerzo (*Reinforcement*)

En esta técnica, el algoritmo o agente interactúa con un ambiente dinámico en el cual debe alcanzar un objetivo específico. Aquí, el algoritmo de aprendizaje recibe premios y castigos a medida que avanza con el problema, y trata de formular por sí mismo la mejor estrategia para obtener el mayor rendimiento sobre el tiempo [39, p. 14].

2.2.2. Clasificación de Algoritmos de tipo Machine Learning

Dentro del aprendizaje automático existen diferentes algoritmos o tipos de modelo que difieren en estructura, el tipo de datos de entrada y salida, y la complejidad computacional.

En esta sección se presentarán los modelos de *Machine Learning* más utilizados para la predicción de crímenes, específicamente para la tarea de clasificación.

Regresión Logística

La regresión logística es un tipo especial de regresión que se utiliza para estimar la probabilidad de que una variable pertenezca a una clase particular. Si la probabilidad estimada es mayor al 50 %, entonces el modelo predice que la variable pertenece a la clase positiva, de lo contrario, se predice que pertenece a la clase negativa [39, p. 144]. La probabilidad estimada se calcula a partir de la siguiente ecuación:

$$\hat{p} = \sigma(X^T \theta) \quad \hat{y} = \begin{cases} 0 & \text{if } \hat{p} < 0,5 \\ 1 & \text{if } \hat{p} \geq 0,5 \end{cases} \quad (2-1)$$

En donde \hat{p} es la probabilidad estimada, σ es la función sigmoide, X es la matriz de características, θ es el vector de parámetros del modelo y \hat{y} corresponde a la predicción.

Máquina de Soporte Vectorial

El modelo de Máquina de Soporte Vectorial, o SVM por sus siglas en inglés, hace parte de los métodos de aprendizaje supervisado para la clasificación, regresión o detección de extremos. En este tipo de modelo se definen funciones base centradas en los puntos de datos de entrenamiento y luego, se selecciona un subconjunto de estos puntos los cuales son llamados vectores de soporte [37]. Las predicciones por SVM son especialmente adecuadas para la clasificación de conjuntos de datos complejos, pero de tamaño pequeño o mediano, además de que pueden realizar tareas de clasificación lineales y no lineales [41] [39, p. 155]. Una propiedad importante de este tipo de modelos es que los parámetros se determinan como la solución a un problema de optimización con función de costo convexa, por lo que, a pesar de que se involucre un problema no lineal, la solución es relativamente sencilla [37, 40].

Árboles de Decisión

Los métodos basados en árboles son conceptualmente simples pero poderosos. La idea principal consiste en seleccionar el modelo de predicción en función de las variables de entrada. Para obtener la estructura del árbol primero se divide el espacio de entradas en un conjunto de rectángulos y luego, a cada región se le asigna un modelo simple como una constante. Luego a esta partición se le asigna su estructura de árbol equivalente en donde cada nodo interno denota una prueba sobre uno o varios atributos, cada rama representa una salida de prueba y los nodos hoja representan las clases [37][39, p. 177]. De esta manera, se obtiene un método de combinación de modelos, en donde un único modelo es responsable de hacer

predicciones para un punto dado del espacio de entradas, y por tanto puede describirse como una secuencia de selecciones binarias correspondiendo a la estructura transversal del árbol [37].

2.2.3. Métricas de Rendimiento

Un paso esencial en todo proyecto de *Machine Learning* es el evaluar el desempeño del modelo de predicción utilizado. Las métricas de evaluación de modelos se utilizan para evaluar el ajuste entre la salida del modelo y los datos. Adicionalmente, permiten la comparación entre diferentes modelos para seleccionar el más adecuado para la tarea o el problema que se esté tratando de resolver [42, 43, 44]. Para las tareas de clasificación se encuentran las siguientes métricas:

Matriz de Confusión

La matriz de confusión se usa para evaluar el rendimiento completo de un clasificador en donde la salida puede ser de dos o más clases. Esta métrica permite mostrar a través de una tabla los tipos de errores que se están cometiendo en el modelo. Su salida corresponde a una matriz de dimensión $N \times N$, en donde N es el número de clases que se están prediciendo y se realiza una comparación entre las predicciones obtenidas y los datos reales. Así, para el caso en que se tienen dos clases A y B , el objetivo principal de la métrica es determinar el número de muestras de la clase A que fueron clasificadas dentro de la clase B [39, p. 93]. En la Figura 2-2 se puede observar la forma estándar de una matriz de confusión para $N = 2$, que se compone de: verdaderos negativos (TN , por sus siglas en inglés), verdaderos positivos (TP , por sus siglas en inglés), falsos negativos (FN , por sus siglas en inglés) y falsos positivos (FP , por sus siglas en inglés). A partir de esta Matriz se pueden construir otras métricas como: Exactitud, Exhaustividad, Precisión y Valor F1 [43].

		Predicciones	
		TN	FP
Datos Reales	TN		
	FN		TP

Figura 2-2: Matriz de Confusión

Exactitud (*Accuracy*)

La exactitud es la métrica más común para evaluar el desempeño de un modelo. Indica el número de elementos clasificados correctamente en comparación con el número total de elementos. Esta métrica tiene limitaciones, ya que no tiene un buen desempeño para conjuntos de datos desequilibrados, es decir, cuando hay una diferencia significativa entre el número de registros con clase *A* o *B*, por lo que la exactitud puede ser más alta de lo que realmente es [44]. Adicionalmente, la interpretación de la exactitud puede ser incorrecta cuando el costo de clasificar erróneamente una de las clases es más perjudicial que la otra. Un ejemplo de ello es el diagnóstico de enfermedades, en donde es más grave diagnosticar incorrectamente a una persona enferma, que realizarle más pruebas a una persona sana [43]. La exactitud puede calcularse a partir de la matriz de confusión como:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2-2)$$

Exhaustividad/Sensibilidad (*Recall*)

Esta métrica indica la tasa de clasificación positiva correcta, es decir, la proporción de positivos que el modelo ha clasificado correctamente en función del número total de muestras positivas [39, p. 94]. Esta métrica tiene como ventaja que no es sensible a los datos desequilibrados.

$$\text{recall} = \frac{TP}{TP + FN} \quad (2-3)$$

Precisión (*Precision*)

La precisión representa la proporción de verdaderos positivos que son correctamente identificados en comparación con el número total de valores positivos que el modelo predijo [39, p. 93]. Una desventaja de la precisión, es que, al igual que la exactitud, es sensible a los datos desequilibrados, es decir, si cambia la proporción de los casos con una de las clases, la precisión informará un valor diferente cuando realmente el desempeño es el mismo [44]. La expresión para la precisión se obtiene como:

$$\text{precision} = \frac{TP}{TP + FP} \quad (2-4)$$

Valor F1 (*F1-Score*)

Esta métrica consiste en una combinación de las métricas de Precisión y Exhaustividad y sirve como un balance entre ellas, y como resultado, esta métrica solo recibe un valor alto si ambos, la Precisión y Exhaustividad tienen valores altos [39, p. 95]. El valor F1 al ser dependiente de la precisión es sensible a los datos desequilibrados entre clases [44].

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (2-5)$$

Por otro lado, para las tareas de regresión se dispone de las siguientes métricas [43]:

Error Cuadrático Medio (MSE)

El MSE, de sus siglas en inglés *Mean Squared Error*, calcula el valor medio de la diferencia al cuadrado entre el valor real y y el predicho \hat{y} para todos los puntos de datos. Todos los valores relacionados se elevan a la segunda potencia, por lo tanto, todos los valores negativos no se compensan con los positivos. Cuanto menor sea el MSE, más precisas serán las predicciones. MSE = 0 es el punto óptimo en el que el pronóstico es perfectamente preciso.

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}^{(i)} - y^{(i)})^2 \quad (2-6)$$

Raíz del Error Cuadrático Medio (RMSE)

El RMSE, de sus siglas en inglés *Root Mean square Error*, es la raíz cuadrada del MSE. Es fácil de interpretar en comparación con el MSE y utiliza valores absolutos más pequeños, lo que es útil para los cálculos informáticos.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}^{(i)} - y^{(i)})^2} \quad (2-7)$$

2.2.4. Métodos de validación

Train-Test Split Evaluation

El método de evaluación *Train-Test Split* es un método que puede utilizarse para problemas de clasificación o regresión y para cualquier algoritmo supervisado. El procedimiento involucra tomar el conjunto de datos y dividirlo entre dos subconjuntos. El primer subconjunto es utilizado para entrenar el modelo y es referido como el conjunto de entrenamiento (*train*). El segundo conjunto se utiliza como insumo para el modelo entrenado y se realizan predicciones para compararlas con los valores esperados. Este segundo subconjunto se refiere como conjunto de evaluación (*test*) [45, 46]. Los porcentajes para el entrenamiento y la evaluación se deben seleccionar de acuerdo con las siguientes consideraciones:

- Costo computacional en el entrenamiento del modelo.
- Costo computacional en la evaluación del modelo.
- Representatividad del conjunto de entrenamiento.
- Representatividad del conjunto de evaluación.

Sin embargo, algunos porcentajes comunes son [45, 46]:

- Entrenamiento: 80 %, Evaluación: 20 %
- Entrenamiento: 70 %, Evaluación: 30 %
- Entrenamiento: 67 %, Evaluación: 33 %
- Entrenamiento: 50 %, Evaluación: 50 %

K-fold Cross-Validation

Es un procedimiento de muestreo que tiene un solo parámetro k que se refiere al número de grupos en los que se debe dividir la información [47]. Es un método popular porque generalmente resulta en menos sesgo. El procedimiento es como sigue:

1. Se separa el conjunto de datos en k grupos.
2. Para cada grupo único:
 - a. Se toma el grupo como grupo de evaluación.
 - b. Se toma el resto de los grupos como conjunto de entrenamiento.
 - c. Se entrena el modelo con el resto de los grupos.
 - d. Se realiza la validación de las métricas con el grupo de evaluación seleccionado.

3. Se realiza un resumen del modelo usando todas las métricas de evaluación.

Nested Cross-Validation

Una de las situaciones comunes que llevan a una evaluación sesgada del desempeño de un modelo es el utilizar el mismo método de validación y el mismo conjunto de datos para sintonizar y seleccionar el modelo. Por ejemplo, el método K-fold Cross-Validation suele utilizarse para la selección de los parámetros del modelo y para estimar el rendimiento del modelo, sin embargo, si el procedimiento se utiliza varias veces con el mismo algoritmo, se puede provocar un sobre ajuste que no permite la generalización del modelo a un nuevo conjunto de datos [48]. Un acercamiento para evitar este sesgo es utilizar la validación cruzada anidada (o del inglés *nested cross-validation*) la cual consiste en anidar un ciclo de validación cruzada tipo k-fold, en el que se ejecuta un problema de optimización para la selección de parámetros, bajo un segundo ciclo de tipo k-fold dedicado el problema de selección del modelo [49]. Dado que se están usando dos ciclos de validación cruzada, este procedimiento también suele llamarse "validación cruzada doble".

2.3. CRISP-DM

En el desarrollo de cualquier tipo de proyectos se requiere de marcos o metodologías que brinden una manera organizada de realizar el trabajo y asignar responsables a las actividades que se van a elaborar.

El modelo de referencia CRISP-DM es el estándar más ampliamente utilizado para el proceso de minería de datos ya que se consolida como una estrategia general de resolución de problemas para cualquier tipo de tema de interés o unidad de estudio. El modelo divide el ciclo de vida de un ejercicio de minería de datos en seis fases diferentes tal y como se visualiza en la Figura 2-3. Estas fases corresponden a la comprensión del tema de interés, la comprensión de los datos, la preparación de los datos, el modelado, la evaluación y el despliegue. A pesar de que hay un camino especificado (ver la dirección de las flechas en la Figura 2-3), la sucesión de las fases no es necesariamente estricta. Adicionalmente, el círculo exterior denota la naturaleza cíclica del proceso, es decir, según las lecciones aprendidas al final del proyecto, se puede decidir volver a repetir las fases anteriores [50]. Cada una de las etapas del modelo CRISP-DM se describe a continuación [2, 51, 52]:

a. Comprensión del tema de interés (*Business Understanding*)

Fase inicial que se enfoca en el entendimiento de los objetivos del proyecto y los requerimientos desde una perspectiva del tema de interés. Se encarga de evaluar la situación actual, traducir los objetivos en términos de un problema de Minería de Datos y de realizar un plan preliminar para el desarrollo del proyecto.

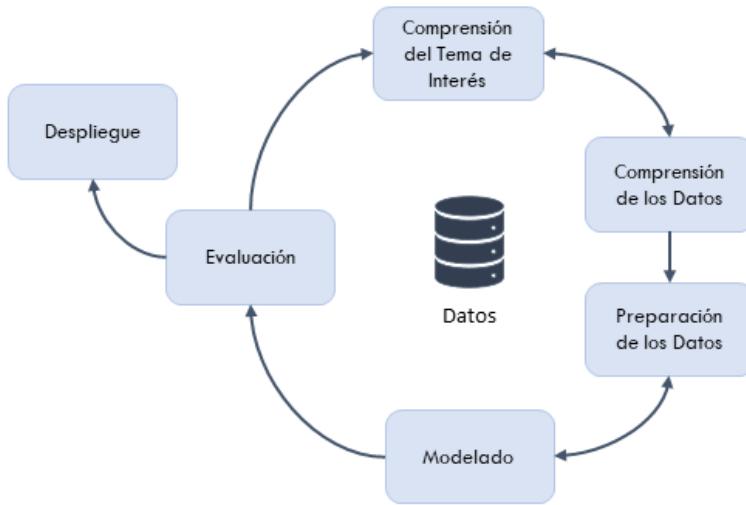


Figura 2-3: Fases de CRISP-DM. Elaborado a partir de [2]

b. Comprensión de los datos (*Data Understanding*)

La fase de comprensión de los datos inicia con una recolección inicial de la información y procede con actividades que permiten familiarizarse con los datos. Se debe realizar un reporte sobre la descripción de los datos, uno sobre la exploración y conocimiento obtenido a partir de los datos y otro sobre la verificación de la calidad de la información. La exploración de los datos puede incluir el resumen de estadísticas, visualizaciones para establecer relaciones y diferentes análisis para la identificación inicial de patrones.

c. Preparación de los datos (*Data Preparation*)

Esta etapa cubre todas las actividades necesarias para la construcción del conjunto de datos final que se utilizará como insumo para los modelos. Es probable que la preparación de los datos se deba realizar varias veces y sin un orden prescrito. Dentro de esta fase se incluye la selección de las tablas, registros y atributos, así como la transformación y la preprocesamiento de los datos.

d. Modelado (*Modeling*)

Una vez se tiene un mayor entendimiento de los datos, se procede al desarrollo y la selección de los modelos de predicción, y se realiza la calibración de los parámetros de los modelos a los valores óptimos. Adicionalmente, se deben plantear las estrategias de validación (incluyendo la división de los datos en los conjuntos de entrenamiento y de prueba) y las métricas de evaluación para medir el desempeño de los modelos. Dependiendo del comportamiento y características del modelo, es posible que se requiera regresar a la fase de preparación de

datos para un mayor refinamiento antes de proseguir con la evaluación.

e. Evaluación (*Evaluation*)

En esta parte del proyecto, se ha construido uno o varios modelos que han utilizado información de alta calidad. Antes de proceder con la última etapa, es importante realizar una evaluación de todos los pasos ejecutados, se realiza una interpretación de los resultados y se verifica que el modelo si alcanza los objetivos planteados. Un punto clave en esta fase es determinar si hay requerimientos que faltan por satisfacer.

f. Despliegue (*Deployment*)

La creación del modelo no es generalmente el final del proyecto. Así el propósito del modelo sea incrementar el entendimiento de los datos, este conocimiento obtenido se necesita organizar y presentar de forma que el usuario lo pueda utilizar. Usualmente esto involucra aplicar modelos en tiempo real dentro de la organización para la toma de decisiones, como por ejemplo, páginas web personalizadas que se encarguen de presentar los resultados. Dependiendo de los requerimientos, la fase de despliegue puede ser tan simple como la generación de un reporte, o tan compleja como la implementación de un proceso repetitivo de Minería de Datos. En muchos casos, es el cliente, y no el analista de datos, el que se encarga de ejecutar el proceso desplegado. Sin embargo, así el analista de datos sea el encargado de ejecutar el proceso desplegado, es importante que el cliente comprenda los procesos operativos que permiten la funcionalidad del proyecto.

Como se ha visto, cada una de las fases puede descomponerse en tareas generales de segundo nivel. En la Figura 2-4 se presenta un esquema de las fases de la metodología CRISP-DM acompañadas por tareas genéricas.



Figura 2-4: Fases de CRISP-DM detallado. Elaborado a partir de [2]

3 Antecedentes y Estado del Arte

A lo largo de la historia, diversas hipótesis se han desarrollado en torno a la teoría del Patrón del Crimen planteada por Brantingham en 1984 [53], en donde se sugiere que el crimen ocurre en áreas determinadas y no es específicamente aleatorio. En particular, se han construido varios modelos de predicción que reúnen información histórica del crimen y la integran con otras variables como factores medioambientales, variables temporales, indicadores demográficos e incluso, con la información publicada en las redes sociales.

En relación con la hipótesis en donde se indica que el calentamiento global puede afectar el aumento de crímenes, en [54] se muestra que para el periodo comprendido entre 1996 y 2013, existe una estrecha correlación entre el número de incidentes y la temperatura ambiente. Específicamente en Finlandia, se obtiene que la variación de la temperatura explica alrededor de un 10 % de la variación del crimen. Asimismo, un estudio realizado entre 1999 y 2004 en Cleveland, Ohio [55], indica que los crímenes más agresivos y las condiciones meteorológicas tienen una gran correlación, cabe resaltar que este resultado se obtuvo luego de aplicar un modelo de regresión lineal entre estas dos variables.

En cuanto a los indicadores demográficos, un estudio realizado en Brasil indica que la tasa de desempleo y los indicadores de analfabetismo son las variables más críticas para describir los homicidios en las ciudades [56]. Como resultado se obtiene que estas dos variables en combinación con el crecimiento de la población masculina explican alrededor del 78 % de la variación del crimen.

Por su parte, para el caso de las variables temporales, en un estudio realizado en Vancouver, Canadá, se puede encontrar que las rutinas de las personas se pueden utilizar para establecer patrones de las actividades criminales [57]. El resultado obtenido muestra que para días específicos de la semana, hay diferentes tipos de crímenes que se presentan con mayor frecuencia.

Finalmente, estudios más recientes combinan información de redes sociales para la construcción de modelos de predicción de actividades criminales. Por ejemplo, en [58] se busca mejorar el desempeño de modelos de predicción basados en el análisis de sentimiento de redes sociales con información meteorológica e histórica de los incidentes.

Por consiguiente, el desarrollo de estas hipótesis ha permitido la configuración de diferentes modelos para la predicción del crimen, siendo la predicción de zonas calientes y la predicción del número de crímenes, las tareas de clasificación y regresión más utilizadas con algoritmos de *Machine Learning*. Teniendo en cuenta que, para que el desarrollo de los algoritmos sea confiable y sus resultados sean adecuados para integrarlos con las estrategias de las autoridades, es necesario realizar una selección adecuada de todos los parámetros que hacen parte de la predicción. Dentro de este conjunto de parámetros se destacan: el área de estudio, el tipo de delito, el rango de tiempo, la división espacial y la división temporal.

Como se puede observar en la tabla 3-1, el planteamiento de algoritmos de *Machine Learning* para la predicción del crimen se ha realizado en países como Estados Unidos, Taiwán, Países Bajos y Brasil, y en relación con el tipo de delito utilizado para la predicción, algunos estudios utilizan toda la información disponible en las bases de datos encontradas, mientras que otros utilizan delitos específicos como el robo residencial. Con respecto a estas bases de datos, se pueden encontrar conjuntos de registros desde el orden de los miles hasta más de un millón, lo que corresponde a un rango de tiempo entre 1 y 40 años para el entrenamiento de los modelos. Por otro lado, en cuanto a las unidades espaciales y temporales, se puede observar que existe una gran variedad de escalas. Respecto a las unidades espaciales, se encuentran estudios que utilizan una malla de celdas de tamaño fijo o variable, mientras que otros utilizan las divisiones administrativas propias de cada región como los barrios. Y en términos temporales, se observa que la predicción puede ir desde 2 semanas hasta un año. Se debe agregar que, en cuanto a los modelos de predicción utilizados, los vecinos cercanos (*k-nearest neighbor*), los bosques aleatorios (*random forest*) y las maquinas de soporte vectorial (*support vector machine*) destacan como los modelos más utilizados para la tarea de clasificación, mientras que para la predicción del número de crímenes se cuenta con los algoritmos de regresión de máquina de soporte vectorial y bosques aleatorios. Para terminar, a modo de comparación, al final de la tabla se pueden observar los parámetros utilizados en la presente tesis.

A pesar de estos resultados, se perciben dos factores que pueden afectar el desarrollo a largo plazo de la investigación en el campo de la predicción del crimen y su implementación en otras regiones del mundo. Por un lado, algunos estudios omiten detalles importantes en los experimentos realizados, y por otra parte, en algunos casos la finalidad de las investigaciones se desvía del objetivo final, el cual consiste en fortalecer las estrategias de las autoridades para la distribución eficiente de los recursos. Por ejemplo, en la ciudad de Bangladesh [59] se realizó una comparación entre varios modelos de regresión de *Machine Learning* para la predicción del número de crímenes en la ciudad. El conjunto de datos que se utiliza para la predicción comprende varios tipos de delitos como robo, secuestro, asesinato y contrabando, y para la predicción se utilizan modelos como regresión lineal y la regresión de árboles aleatorios. Sin embargo, a diferencia de otros artículos, los autores no realizan una división espacial

Área de Estudio	Rango de Tiempo	Tipo de Delito	Número de Muestras	Tipo de Inferencia	Tarea	Unidad Espacial	Unidad Temporal	Modelos	Métricas	Referencia
Nueva York, USA	Enero 2014 - Abril 2015	7 tipos de crimen	ND	Zonas Calientes	Clasificación Binaria	Malla de celdas de 0,01 Latitud x 0,01 Longitud	Da, semana, mes	Logistic Regression, Naive Bayes, Support Vector Machine, Neural Network, Decision Tree	Accuracy, F1 score, AUC	Yang et al. (2018)
USA	ND	Robo residencial	ND	Zonas Calientes	Clasificación Binaria	Malla de celdas (ND)	Mes	Decision Tree, Support Vector Machine, Naive Bayes	Accuracy, F1-Score	Yu et al. (2011)
Portland, USA	Marzo 2012 - Diciembre 2016	Todos los Crímenes	ND	Zonas Calientes	Clasificación Binaria	Malla de celdas de 183 m x 183 m	2 Semanas	Spatio-Temporal Neural Network (STNN)	Accuracy, Precision, Recall, F1-score	Zhuang et al. (2017)
Natal, Brasil	2006-2016	ND	ND	Zonas Calientes	Clasificación Binaria	Celdas de tamaño variable	Semana	K-Nearest Neighbor, MLP, Random Forest	Accuracy	Araújo et al. (2018)
Taoyuan City, Taiwan	2015-2018	Robo de Autos	≈ 8580	Zonas Calientes	Clasificación Binaria	Malla de celdas de 5 a 100 x 5 a 100	Mes	K-Nearest Neighbor, Support Vector Machine, Random Forest	Accuracy, Precision, Recall, F1-score	Lin et al. (2018)
Ámsterdam	2011-2014	3 tipos de crímenes	163,800	Zonas Calientes	Clasificación Binaria	Malla de celdas 200 m x 200 m	2 semanas, mes	Logistic Regression, Neural Network, Ensemble Model	Precision	Rummens et al. (2017)
Chicago, USA	2011-2015	34 tipos de crímenes	6,6 millones	Número de Crímenes	Regresión	Comunidades	Mes, año	Polynomial Regression, Support Vector Regression, Auto-regressive model	Precision	Dash et al. (2018)
Nueva York, USA	2014-2015	Todos los crímenes y 5 tipos de crimen	174,682	Número de Crímenes	Regresión	Bloque de Censo	Año	Random Forest regressor, Extra-Tree Regressor, Gradient-Boosting Regressor	MSE	Kadar and Pietikosa (2018)
USA	1960-2009	2 tipos de crimen	ND	Tasa de crímenes	Regresión	Región	Año	Naïve Models	RMSE	Shoemaker (2013)
San Francisco, Natal	2003-2015 2006-2016	Crímenes Violentos	ND	Número de Crímenes	Regresión	Cluster de K-means	Semana	Support Vector Regression, Multi Layer Perceptron Regression, Random Forest Regression	MSE	J. Borges (2018)
Medellín	Enero 2015 - Diciembre 2019	Atraco, Desuelto, Cosquilleo, Rapapolazo	215,000	Zonas Calientes	Clasificación Binaria	Malla de Celdas 610 m x 610 m	2 Semanas	Logistic Regression, Random Forest, Support Vector Machine	Accuracy, Precision, Recall, F1-score	Munoz, Victor (2021)

Tabla 3-1: Compración de los estudios realizado en torno a la predicción del crimen (ND = No Definido) [5, 6, 7, 8, 9, 10, 11, 12, 13, 14].

de la ciudad, y la ventana temporal para la predicción es de un año. Y aunque se menciona que estos resultados podrían ayudar a las autoridades para la toma de medidas preventivas, se percibe que este tipo de acercamientos no son suficientes para la definición de estrategias efectivas si se quiere realizar una distribución inteligente del equipo de la policía en la región.

Por su parte, en la ciudad de Ámsterdam se realizó un estudio para la predicción del crimen utilizando un método de ensamble compuesto por una regresión logística y un modelo de redes neuronales [9]. Para su ejecución, se utilizó dentro del conjunto de datos de entrenamiento, información sobre los eventos históricos de crimen, variables demográficas, socio económicas y ambientales. Estos datos se utilizan con el objetivo de predecir el riesgo de crimen a nivel de una cuadrícula, en una ventana quincenal, y se realiza una comparación con una predicción mensual desagregada (predicciones diurnas y nocturnas). Con los resultados obtenidos, el autor expresa la viabilidad de utilizar este tipo de modelos para colaborar con las autoridades, e indica que la división temporal tiene un impacto importante en el desempeño de los modelos, ya que las predicciones desagregadas obtuvieron una mayor puntuación en las métricas de evaluación implementadas.

Adicionalmente, en Londres se obtuvo una precisión de casi el 70 % al predecir si un área específica de la ciudad puede ser una zona caliente [60]. Para ello, los autores integraron información del crimen, con los datos capturados por la infraestructura de la red móvil de la ciudad, y evaluaron cinco clasificadores de *Machine Learning*: regresión logística, máquina de soporte vectorial, redes neuronales, árboles de decisión y diferentes métodos de ensamble. No obstante, se puede observar que el estudio estuvo limitado por los pocos datos históricos de la red móvil y la información de los incidentes estaba agrupada por meses. Cabe señalar que, al igual que la división espacial, la división temporal es un factor importante para el desarrollo de las estrategias que permitan mejorar los tiempos de respuesta por parte de las autoridades, y esta división debe estar alineada con las necesidades de la policía, es decir, si se está buscando una solución de actuación inmediata o de planeación estratégica (corto o largo plazo).

En la ciudad de Portland, [7] utiliza la base de datos de los registros de las llamadas realizadas a la policía, para crear una red neuronal recurrente con una ventana temporal de dos semanas. Las tres arquitecturas evaluadas de redes recurrentes obtuvieron un desempeño de 81,5 % en la exactitud, una precisión del 86-87 %, una exhaustividad de 75 % y un valor F1 del 80 %. Sin embargo, solo están considerando información sobre el crimen, no tienen información sobre la semana, la hora, el clima ni indicadores demográficos.

Por último, en [10], se utiliza una base de datos de robo de automóviles para diferentes modelos de *Machine Learning* para la predicción de crímenes en Taoyuan, Taiwan. Se evalúan modelos de clasificación como: bosques aleatorios, máquina de soporte vectorial, vecinos más

cercanos y redes neuronales profundas. Siendo este último modelo, el de mejor desempeño con un valor F1 de 47,34 %, una exactitud de 83,76 %, una precisión de 41,07 % y una exhaustividad de 57,08 %. En este estudio los autores incluyen dentro de las características información sobre el número de crímenes que ocurrieron en la división espacial para los meses anteriores y los lugares de interés cercanos al incidente como: supermercados, peluquerías, farmacias, etc. Sin embargo, la ventana de tiempo seleccionada es muy ancha, 1 mes, comparada con otros modelos, lo que puede afectar el desempeño y en consecuencia las decisiones que puedan tomar las autoridades.

Cabe resaltar que estos resultados son el producto del trabajo realizado por investigadores en diferentes ramas del conocimiento, y cada uno de sus aportes se pueden utilizar para la construcción de una estrategia replicable a diferentes necesidades. La presente tesis pretende evaluar modelos de *Machine Learning* para la predicción del crimen específicamente en la ciudad de Medellín, mediante el desarrollo de una metodología que se pueda adaptar a las necesidades de las autoridades y se pueda replicar, no solo a nivel nacional sino internacionalmente. Y en consecuencia, se espera que los resultados se utilicen como un insumo para soportar plataformas inteligentes para la gestión eficiente de los recursos de las autoridades.

A nivel internacional se puede observar que las autoridades de las principales ciudades han reforzado el registro de actividades delictivas y han puesto al servicio de los ciudadanos diversas fuentes de información para el desarrollo de estrategias que permitan mejorar la respuesta ante incidentes. Por ejemplo, en la ciudad de Natal (Brasil) [61] las autoridades cuentan con una aplicación web que les ayuda a pronosticar la incidencia de delitos. Esta información les permite elaborar una lista de ubicaciones predefinidas y de tiempos de patrullaje para cada vehículo policial. Para obtener estos resultados, la aplicación utiliza los registros de incidentes históricos e implementa múltiples métodos de aprendizaje estadístico y de *Machine Learning*. Sin embargo, se puede observar que esta estrategia no explora más variables que puedan ayudar a reforzar los modelos, como las variables demográficas o socioeconómicas, y en cuanto a la división espacial, los diseñadores se ven limitados por los cuadrantes predefinidos por la policía. Esto último representa un punto de discusión para el desarrollo de nuevas estrategias, ya que, si bien la variación de las divisiones espaciales para la predicción puede ayudar al desempeño de los modelos desarrollados [15], siempre se deben alinear estas decisiones con las estrategias de la policía.

A nivel nacional, en la ciudad de Bogotá se encuentra en desarrollo el proyecto “Diseño y Validación de Modelos de Analítica Predictiva de Fenómenos de Seguridad y Convivencia para la Toma de Decisiones en la ciudad de Bogotá” [62]. Este proyecto de investigación y desarrollo experimental inició su ejecución en julio del 2019 y se encuentra conformado por el Sistema General de Regalías, la Universidad Nacional de Colombia - sede Bogotá, la empresa Quantil y la Secretaría Distrital de Seguridad, Convivencia y Justicia. El proyecto

busca mejorar la gestión de los datos sobre criminalidad y formular métodos basados en analítica predictiva adaptados a las condiciones sociales particulares de Bogotá, para mejorar el entendimiento y predictibilidad de los fenómenos de seguridad y convivencia sobre el conjunto de datos disponibles.

En la ciudad de Medellín, la empresa EMP ha creado un Sistema de Información para el Análisis del Entorno de EPM a partir de los datos de MEData y otras fuentes [63]. Este sistema busca incorporar herramientas de analítica, con un enfoque geográfico, para agilizar la respuesta a las dinámicas de los territorios, y en la entrega de señales oportunas en materia de seguridad. Con este ejercicio, y con la ayuda del Sistema de Información para la Seguridad y Convivencia de la Alcaldía, han logrado alimentar Dashboards y Mapas de calor que les permite conocer la situación de seguridad en el territorio de forma objetiva y desde varios niveles espaciales (comuna, barrio). Actualmente el sistema contiene información de criminalidad, social, política y económica y de inversión en la ciudad de Medellín, pero se encuentran trabajando en ampliar el análisis a toda el Área Metropolitana del Valle de Aburrá y a nivel nacional.

A partir de todas las estrategias abordadas, se ve la oportunidad de proponer para la ciudad de Medellín, un modelo de *Machine Learning* que integre la información histórica de crímenes con diversas fuentes de información, tales como variables demográficas, socioeconómicas y ambientales. Teniendo en cuenta los desafíos que presentan este tipo de soluciones, en cuanto a la definición de las divisiones temporales y espaciales, la disponibilidad de la información y la calidad de los datos. Así mismo, no se debe perder el objetivo, que consiste en ayudar a las autoridades, por lo que se deben alienar todos los parámetros de diseño con sus estrategias.

Esta tesis se está realizando bajo el proyecto “Gestión inteligente de problemas de seguridad ciudadana a través de modelos y herramientas generadas a partir de plataformas para territorios inteligentes apoyadas por estrategias de participación ciudadana” liderado por una alianza entre la Universidad de Antioquia y la Universidad Nacional de Colombia - Sede Medellín y con enfoque en el Área Metropolitana del Valle de Aburrá. En general el proyecto busca el desarrollo de estrategias para la prevención del riesgo y el fortalecimiento de la seguridad ciudadana, mediante el uso de tecnologías de percepción electrónica e inteligencia computacional, que ayuden a los organismos de seguridad en la toma de decisiones. Esta tesis está dirigida a tener un primer acercamiento a las fuentes de información disponibles en la ciudad, la caracterización del crimen y de las estrategias actuales de las fuerzas de seguridad, y en la investigación de modelos de *Machine Learning* que apoyen las estrategias para la prevención del crimen.

4 Metodología para la Predicción de Crímenes

De acuerdo con la literatura, se puede observar que existen múltiples técnicas para la predicción de los crímenes en la ciudad. Desde el punto de vista de los modelos utilizados, se pueden encontrar modelos auto regresivos, modelos de aprendizaje automático y hasta algoritmos de aprendizaje profundo. Por parte de las inferencias, diversos estudios han explorado la predicción de zonas calientes en una ciudad, la predicción del número y la tasa de crímenes en el tiempo. Todas estas posturas buscan brindar herramientas a las autoridades para tener una mayor cobertura de la distribución de los delitos a corto y largo plazo en la región, y con esto, mejorar sus estrategias para la gestión de los recursos y para atender de manera oportuna los incidentes.

Aunque estos estudios han mostrado resultados concretos, se puede encontrar que algunos investigadores no reportan adecuadamente todos los detalles necesarios para la reproducción de los resultados. Uno de los factores que influye en este comportamiento, son los diversos antecedentes de los investigadores (criminología, ciencia de los datos, autoridades, ciencias sociales, etc.). Como se pudo observar en la Tabla 3-1, y según el trabajo realizado en [15], los proyectos de investigación dentro del campo deberían reportar como mínimo los parámetros descritos en la Tabla 4-1.

Tabla 4-1: Parámetros que se deben reportar en la predicción del crimen [15].

ID	Parámetro	Ejemplo
1	Área de estudio	Medellín, Nueva York, Brasil
2	Escala	Ciudad, Departamento, País,
3	Periodo de tiempo	2004-2014
4	Meses	30
5	Tipo de crimen	Todos los crímenes, Hurto, Robo
6	Tamaño de la muestra	310000
7	Inferencia	Hotspot, # de crímenes, tasa del crimen
8	Tarea	Clasificación, Regresión
9	Escala Espacial	1 año, semana, día
10	Escala Temporal	División Administrativa, Malla Espacial

Con el objetivo de resolver esta necesidad, la presente tesis de maestría propone la integración de los parámetros descritos en la Tabla 4-1 con la metodología *CRISP-DM* para la Predicción del Crimen con Modelos de *Machine Learning*. El objetivo de esta capítulo consiste en revisar cada una de las tareas de *CRISP-DM*, y enfatizar en las actividades particulares que se deben llevar a cabo para la predicción del crimen.

4.1. Desarrollo de la Metodología

A continuación se presenta el detalle de las actividades que componen la metodología, especificando los resultados que se deberían entregar al término de cada una de las tareas.

Tabla 4-2: Metodología para la predicción del crimen con modelos de Machine Learning

ID	Fase	Tarea	Resultado
1	Comprensión del Tema de Interés	Determinar los Objetivos del Tema de Interés	<ul style="list-style-type: none"> • Patrocinador/Beneficiarios • Objetivos • Área de Estudio (Escala)*
		Evaluación de la Situación	<ul style="list-style-type: none"> • Inventario de Recursos
		Determinar los Objetivos de Minería de Datos	<ul style="list-style-type: none"> • Objetivos de Aprendizaje Automático • Tipo de Inferencia*
		Generar un Plan de Proyecto	<ul style="list-style-type: none"> • Lista de Actividades
2	Comprensión de los Datos	Recopilación de los Datos	<ul style="list-style-type: none"> • Reporte de la Recopilación
		Descripción de los Datos	<ul style="list-style-type: none"> • Reporte de la Descripción
		Exploración de los Datos	<ul style="list-style-type: none"> • Reporte de la Exploración
		Verificación de la Calidad de los Datos	<ul style="list-style-type: none"> • Reporte de la Calidad de Datos
3	Preparación de los Datos	Selección de los Datos	<ul style="list-style-type: none"> • Motivos de Inclusión y Exclusión • Periodo de Tiempo (Meses)* • Tipo de Crimen* • Variables
		Preprocesamiento	<ul style="list-style-type: none"> • Reporte del Preprocesamiento
		Construcción de los Datos	<ul style="list-style-type: none"> • Lista de Atributos Generados • Lista de Registros Generados
		Integración de los Datos	<ul style="list-style-type: none"> • Lista de Integraciones Realizadas
		Formateo de los Datos	<ul style="list-style-type: none"> • Formato de los Datos
			<ul style="list-style-type: none"> • Conjunto de Datos Final • Tamaño de la Muestra* • Escala Espacial* • Escala Temporal*
4	Modelado	Selección de la Técnica de Modelado	<ul style="list-style-type: none"> • Tarea de Predicción* • Modelo Base* • Lista de Modelos*
		Diseño de la Evaluación	<ul style="list-style-type: none"> • Estrategia de Validación* • Métricas*
		Construcción del Modelo	<ul style="list-style-type: none"> • Configuración de Hiperparámetros
		Evaluación del Modelo	<ul style="list-style-type: none"> • Resultados del Modelo
5	Evaluación	Evaluación de Resultados	<ul style="list-style-type: none"> • Mejor Modelo*
		Revisar el Proceso	
		Establecimiento de los Siguientes Pasos	

1. Comprensión del tema de interés

En términos de un proyecto que busca realizar una predicción del crimen, esta primera etapa se debe enfocar en el entendimiento de los objetivos de la línea de investigación y los requerimientos de las personas que se encuentran apoyando el proyecto. Se deben plantear los objetivos de la predicción en términos de un objetivo de Aprendizaje Automático, es decir, se debe definir la tarea de predicción (clasificación, agrupamiento, predicción, etc.) y cuáles son las características que se esperan en el conjunto de datos. Finalmente se debe realizar un plan preliminar orientado al cumplimiento de los objetivos.

- a. **Determinar los objetivos del tema de interés:** En este punto se debe identificar el patrocinador del proyecto o a las personas que se pueden ver beneficiadas con los resultados del trabajo de investigación. Se debe realizar una definición de los objetivos, en términos del área de estudio (ciudad, departamento, país), el tipo de predicción (corto plazo o largo plazo), y los resultados esperados. **Resultado:** patrocinador/beneficiado, objetivos, área de Estudio.
- b. **Evaluación de la situación:** Esta tarea involucra un análisis detallado de los recursos disponibles, los supuestos, y otros factores que son determinantes para alcanzar los objetivos y para la ejecución del proyecto. En este punto se debe crear una lista de todas las fuentes de información disponibles, los recursos computacionales y el personal con el que se va a contar para el desarrollo del proyecto. **Resultado:** Inventario de recursos.
- c. **Determinar los objetivos de minería de datos:** En este punto se deben plantear los objetivos del proyecto en términos del aprendizaje automático. Para lograrlo, se debe determinar el tipo de inferencia sobre la predicción del crimen en la que se desea trabajar, como por ejemplo:
 - Predicción de zonas calientes
 - Predicción de número de crímenes
 - Predicción de la tasa del crimen
 - Predicción de la categoría del crimen
 - Encontrar las propiedades de un clúster.
 - Encontrar el porcentaje de crimen en un clúster.**Resultado:** Objetivos de aprendizaje automático, tipo de inferencia.
- d. **Generar un plan de proyecto:** Para finalizar esta primera fase, se deben enumerar las actividades que se tienen contempladas para obtener los resultados. **Resultado:** Lista de actividades.

2. Comprensión de los Datos

En esta fase se debe realizar una enumeración de las fuentes de información disponibles, relacionadas con el registro de crímenes en la región de estudio, y aquellas que puedan utilizarse para la predicción del crimen. Se debe realizar un reporte sobre la descripción de las fuentes, uno sobre el conocimiento obtenido luego de realizar un proceso de exploración y otro con la identificación de problemas de calidad de la información.

- a. **Recopilación de los datos:** En este punto se realiza la recolección de los datos que se listaron en los recursos del proyecto. Se deben listar todos los conjuntos de datos, la ubicación, el formato, los métodos utilizados para adquirirlos, y los problemas encontrados durante la consulta. Se deben registrar los problemas encontrados y las posibles soluciones que se implementaron. **Resultado:** Reporte de la recopilación.
- b. **Descripción de los datos:** En esta parte se deben examinar las propiedades de los conjuntos de datos adquiridos. Se debe incluir el formato de la información, la cantidad de información (por ejemplo el número de registros en la tabla) y las características. Asimismo, se debe evaluar si la información adquirida puede servir para los requerimientos del proyecto. **Resultado:** Reporte de la descripción.
- c. **Exploración de los datos:** En esta parte se utilizan técnicas de consulta, visualización y reporte de información. En este punto se incluye el reporte de las distribuciones de parámetros claves, la relaciones entre dos o más variables, agregaciones simples y análisis estadísticos. **Resultado:** Reporte de la exploración.
- d. **Verificación de la calidad de los datos:** En este punto se examina la calidad de la información, se deben registrar los resultados de las verificaciones que se realizan con respecto a la validación de los datos. **Resultado:** Reporte de la calidad de los datos.

3. Preparación de los Datos

En esta fase se deben cubrir todas las actividades necesarias para la construcción del conjunto de datos final. El proceso inicia con la selección de los datos, en donde se deberán incluir la descripción de todas las decisiones tomadas, y para el objeto de estudio actual, se deberá especificar el tipo de crimen sobre el cual se está realizando la predicción. Es probable que la preparación de los datos se deba realizar varias veces, sin seguir un orden prescrito, sin embargo, es necesario dejar una documentación sobre las tareas de construcción de datos, como la generación de mallas espaciales o ventanas temporales, así como la documentación de la información de los crímenes con fuentes de información meteorológicas o demográficas. Finalmente, se debe presentar un resumen del conjunto de datos definitivo, en donde se indique el tamaño de la muestra y en caso de aplicar, la escala temporal y espacial.

- a. **Selección de los datos:** En este punto se deben decidir cuáles son los conjuntos de datos que se van a utilizar para la predicción del crimen. Los criterios que se pueden utilizar son: la relevancia de la información para obtener los objetivos, la calidad, las limitaciones en el volumen y el tipo de dato. Se debe registrar la información que se incluye y la que se excluye y los motivos de la decisión. Es necesario dejar descrito el tipo de crimen con el que se va a trabajar (hurto, robo, robo de automóviles, robo de propiedad, homicidios, etc.), el periodo de tiempo con el que se va a trabajar, y las variables dependientes o independientes, adicionales a las fuentes de información de crimen. **Resultado:** Motivos de inclusión y exclusión, periodo de tiempo, tipo de crimen, variables.
- b. **Preprocesamiento de datos:** A partir de los reportes de calidad de la información y las decisiones tomadas en el numeral anterior, se deben aplicar técnicas como la selección de datos, inserción de datos apropiados, o técnicas más avanzadas como la de estimación de datos ausentes. Se debe registrar el número de registros finales después del preprocesamiento. **Resultado:** Reporte del preprocesamiento.
- c. **Construcción de los datos:** En este punto se debe realizar la producción de atributos derivados de otros atributos, o la transformación de nuevos atributos. **Resultado:** Lista de atributos generados, Lista de registros generados.
- d. **Integración de los datos:** Dado que se pueden utilizar diversas hipótesis para la predicción de los crímenes, se ve la necesidad de integrar la información de los incidentes con los datos de otras fuentes de información, como lo pueden ser: bases de datos meteorológicas o demográficas. Se debe realizar una descripción de los procesos de integración utilizados, como la agregación de tablas. **Resultado:** Lista de integraciones realizadas.
- e. **Formateo de los datos:** Finalmente, con el propósito de evitar contratiempos en la programación, se debe asegurar la correcta asignación del tipo de formato a todos los atributos. **Resultado:** Formato de los datos.

Finalmente, esta fase debe contar con los siguientes productos antes de continuar con la fase de modelado: conjunto de datos final, tamaño de la muestra, escala espacial y escala temporal.

4. Modelado

En esta fase se debe realizar la selección de los modelos de predicción tomando como referencia el tipo de inferencia seleccionada en la fase de comprensión del tema de interés, y el tipo de tarea que se desea realizar con el conjunto de datos definitivo (clasificación, regresión, agrupación). Se debe establecer la estrategia de validación de los modelos y las métricas de evaluación que permitirán comparar los desempeños de los diferentes modelos.

- a. **Selección de la técnica de modelado:** Dependiendo del tipo de inferencia seleccionada en los objetivos y el conjunto de datos definitivo, se debe seleccionar la técnica de modelado que se va a utilizar. Para esta metodología se propone contar con un modelo base y varios modelos propuestos para la comparación. **Resultado:** tarea de predicción, modelo base, lista de modelos.
- b. **Diseño de la evaluación:** Antes de proceder con la construcción del modelo, se debe generar un mecanismo para evaluar la calidad del desempeño. Se deben definir las métricas de evaluación que se van a utilizar y la estrategia de validación (train-test split, validación cruzada). **Resultado:** Estrategia de validación, métricas.
- c. **Construcción del modelo:** Se deben utilizar las herramientas definidas en los recursos del proyecto, para implementar los modelos propuestos. Específicamente, se deben calibrar y reportar los hiperparámetros óptimos para cada uno de los modelos. **Resultado:** Configuración de hiperparámetros.
- d. **Evaluación del modelo:** Se deben listar los resultados de las métricas para cada uno de los modelos utilizados. **Resultado:** Resultados del modelo.

5. Evaluación

En esta última fase, se deben revisar los resultados obtenidos en el modelado y verificar si se cumplen con los objetivos planteados. Un punto clave en esta fase es determinar si hay requerimientos que faltan por satisfacer.

- a. **Evaluación de resultados:** En este punto se debe revisar el grado en el que el modelo se adhiere a los objetivos de las autoridades y determinar si es suficiente o si se deben realizar modificaciones. **Resultado:** mejor modelo.
- b. **Revisar el proceso:** Se debe realizar un análisis de todas las actividades ejecutadas y determinar si se han reportado todos los detalles expuestos en la presente metodología.
- c. **Establecimiento de los siguientes pasos:** Dependiendo del análisis realizado en el punto anterior, se debe determinar si se deben ejecutar actividades adicionales, si se puede terminar con el proyecto o si se debe continuar con la fase de despliegue.

Como se puede observar en el capítulo 2, la metodología CRISP-DM incluye una fase adicional que consiste en el despliegue de la solución, sin embargo, se espera que los resultados de esta tesis se puedan utilizar para soportar plataformas inteligentes en una fase posterior del proyecto macro en el que se encuentra inscrita la investigación (Minciencias BPIN-2020000100044).

5 Análisis del Problema y Recolección de Datos

Con el objetivo de validar la metodología planteada en la presente tesis, se propone realizar el desarrollo de cada una de las actividades para la predicción de delitos en la ciudad de Medellín. Por lo tanto, en este capítulo se realiza una identificación de las fuentes de información disponibles que contribuyan con el desarrollo de los objetivos. Se presenta un análisis descriptivo y exploratorio de los datos seleccionados y se finaliza con los procesos de integración, construcción y preprocesamiento de los datos.

5.1. Comprensión del tema de interés

El propósito de este trabajo es proponer y evaluar un modelo para la predicción de crímenes en la ciudad de Medellín que permita a las autoridades mejorar la gestión de sus recursos y atender de manera apropiada a los incidentes que se presentan en la ciudad. Dados los indicadores sobre criminalidad revisados en la introducción, se toma la decisión de trabajar con el hurto a persona, ya que esta categoría es la que presenta las mayores tasas [1].

Actualmente, la alcaldía de Medellín cuenta con una base de datos de los hurtos reportados entre enero de 2003 y septiembre del 2020. Adicionalmente, otras entidades del gobierno como el DANE y SIATA almacenan información pública sobre variables meteorológicas e indicadores demográficos que podrían aportar a los modelos de predicción. Por otro lado, en cuanto a las herramientas disponibles para la construcción y evaluación de los modelos, se cuenta con la plataforma de *Google Colab*, el lenguaje de programación *Python* y la librería de *Scikit-Learn*, que contiene los algoritmos de *Machine Learning*, las estrategias de validación y las métricas de evaluación de desempeño.

Desde un punto de vista más técnico, se espera encontrar un modelo de *Machine Learning* que permita la predicción de zonas calientes en la ciudad. Para este caso, se propone un conjunto de datos en donde cada uno de los registros contiene información correspondiente a una división espaciotemporal en la ciudad de Medellín. El conjunto de características o atributos, puede contener información histórica de los crímenes, variables meteorológicas e indicadores demográficos.

Finalmente, para proceder con la construcción y validación de los modelos de *Machine Learning*, se decide tomar como referencia las actividades reportadas en [64]:

- a. Definir una división espacial sobre el mapa geográfico de la región seleccionada y una división temporal o ventana de tiempo sobre el periodo de tiempo seleccionado.
- c. Distribuir los registros de los delitos en cada una de las divisiones espaciales, y separarlos por cada una de las ventanas temporales.
- d. Definir las zonas calientes para cada una de las ventanas temporales.
- e. Las unidades espaciales que no son clasificadas como zona caliente para alguna de las ventanas temporales son descartadas.
- f. Se procede con la integración con otras fuentes de información que puedan aportar a la tarea de predicción.
- g. Se procede con la construcción y evaluación de los modelos.

5.2. Estudio y Comprensión de los Datos

A continuación, se presenta la lista de las fuentes de información disponibles en la ciudad de Medellín y que son relevantes para la presente investigación. También, se realiza la descripción y exploración de los datos y se hace una verificación sobre la calidad de la información.

5.2.1. Fuentes de Información

Actualmente la ciudad de Medellín cuenta con diversas fuentes de información, entre las cuales se pueden encontrar bases de datos con los siguientes aspectos:

- **Base de datos de crímenes y accidentes**

La página de MEDATA de la alcaldía de Medellín, reúne información sobre varios aspectos sociales en la ciudad como salud, seguridad, educación, cultura, movilidad y población [65]. Específicamente, este trabajo se enfoca en las bases de datos del componente de seguridad, y como se podrá observar en la Preparación de los Datos, los modelos de predicción se entrenarán con información de hurtos a personas en las modalidades de atraco, descuido, cosquilleo y raponazo.

Cada una de las bases de datos sigue el modelo unificado de seguridad y convivencia de la ciudad que facilita mantener de una forma estructurada todos los hechos relacionados con la seguridad, la convivencia, derechos humanos y justicia que han ocurrido en la ciudad de Medellín. Para obtener la información, solo se debe ingresar al sitio web

<http://metadata.gov.co/dataset/hurto-persona>, y utilizar la opción de descarga como se muestra en la Figura 5-1. Una vez realizado el proceso, el sitio provee un archivo en formato *csv* que se puede leer directamente en Python mediante la librería de Pandas [66].



Figura 5-1: Recopilación de los registros de hurto a persona.

■ Base de datos de las comunas y barrios de Medellín

Por otro lado, la alcaldía también cuenta con otro portal llamado GEO MEDELLIN en donde se dispone de información acerca de la infraestructura, catastro y ordenamiento territorial [67]. Específicamente se puede acceder a información como la malla vial de la ciudad, la distribución del código postal en la región, el espacio público existente y la distribución de los barrios y comunas. Para este trabajo de investigación, se utiliza el archivo en formato *Shapefile* que contiene los polígonos de las superficies de los barrios y las comunas de la ciudad. De igual forma, esta información se descarga directamente del portal (Figura 5-2), y por medio de *Python*, mediante la librería de GeoPandas, se puede realizar la consulta, visualización y modificación de los diferentes sectores espaciales.

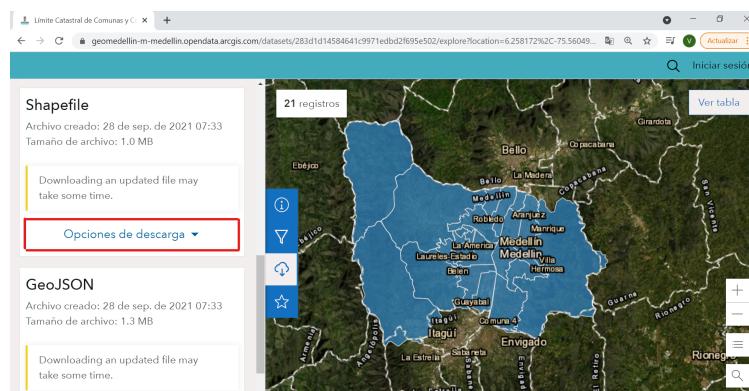


Figura 5-2: Recopilación de los polígonos de las comunas de Medellín.

■ Base de datos de las variables meteorológicas

Para este punto se encuentra que el SIATA (Sistema de Alertas Temprana de Medellín y el Valle de Aburra), tiene datos disponibles acerca de varias variables meteorológicas en la ciudad como: temperatura, precipitación, presión, viento, radiación. Así mismo, esta entidad monitorea la calidad del aire en la ciudad, y cuenta con algunas estaciones hidrológicas para medir el nivel y velocidad del río de la ciudad.

Para consultar la información, se debe realizar el registro en la plataforma para obtener credenciales de usuario y contraseña. En la Figura 5-3 se puede observar la ventana de inicio al sistema. Una vez se realiza el ingreso a la plataforma, se debe indicar la variable que se desea consultar, un rango de fechas y las estaciones que almacenan la información. Este paso se puede evidenciar en la Figura 5-4. Finalmente, la información se puede descargar en formato *csv* a través de los enlaces dispuestos. Cada archivo contiene información sobre un solo mes y para una sola estación, sin embargo, la lectura e integración de los diferentes archivos se puede realizar por medio de la librería Pandas.



Figura 5-3: Ingreso al portal SIATA.



Figura 5-4: Consulta de variables meteorológicas SIATA.

■ Base de datos de las variables demográficas

En este punto se encuentra al Departamento Administrativo Nacional de Estadísticas (DANE) como la fuente de información que contiene bases de datos sobre el crecimiento de la población y algunas métricas como la tasa de desempleo. Aunque varias de sus métricas se realizan a nivel nacional, el DANE también pública información sobre variables de las principales ciudades del país en donde se encuentra Medellín [68].

Para realizar la consulta de la información, se ingresa al portal en la sección de documentos históricos y se selecciona el enlace con el último reporte, tal y como se muestra en la Figura 5-5. Para este caso, la información se reporta en formato *xlsx* y se puede leer por medio de la librería de Pandas.

-GEIH- 2021				
Periodo	Documento	Documento	Documento	Anexos
Agosto	Boletín técnico	Comunicado de prensa	• Presentación (rueda de prensa) • Presentación (extendida)	• Anexos • Anexos desestacionalizadas

Figura 5-5: Recopilación de indicadores demográficos DANE.

5.2.2. Descripción y Exploración de los Datos

En línea con el objetivo planteado en la comprensión del tema de interés, en este punto se procede con la descripción y exploración de las bases de datos que contienen los registros sobre los hurtos a personas, la temperatura y el desempleo en la ciudad de Medellín.

Base de hurtos a personas

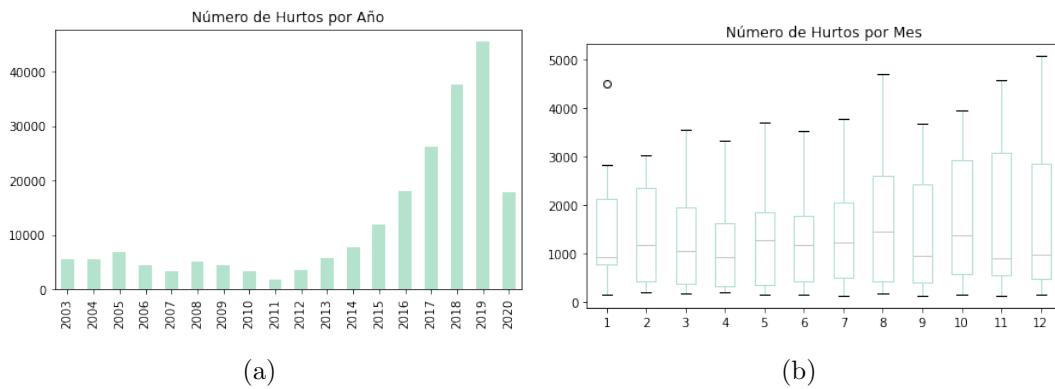
Esta base de datos se compone de 215.279 registros y 35 campos que se pueden organizar en 6 categorías como se muestra en la Tabla 5-1. En particular, cabe restaltar que la entidad solamente disponibiliza los registros hasta un año antes de la fecha de consulta.

■ Información temporal

Para la fecha en la que se realizó la consulta de la presente investigación, la información de incidentes se encontraba desde el 01 de Enero de 2003, hasta el 04 de septiembre de 2020. Asimismo, como se muestra en la Figura 5-6, los años con el mayor porcentaje de registros corresponden al 2019 (21.13 %), 2018 (17.49 %) y 2017 (12.21 %) abarcando un 50.82 % de toda la muestra. Adicionalmente, en la escala de meses se puede observar una gran dispersión, siendo agosto el mes con la mayor mediana.

Tabla 5-1: Campos que componen la base de datos de hurtos

Categoría	Campos
Temporal	Fecha del hecho
Espacial	Latitud, longitud, barrio, nombre del barrio, código de barrio, código comuna, lugar, sede receptora
Características demográficas de la víctima	Sexo, edad, estado civil
Información del hurto	Cantidad, medio de transporte, modalidad, conducta especial, arma medio, conducta
Características del bien robado	Bien, categoría bien, grupo bien, modelo, color
Sin Información	Grupo actor, actividad delictiva, parentesco, ocupación, discapacidad, grupo especial, nivel académico, testigo, caracterización, artículo penal, categoría penal, permiso

**Figura 5-6:** Comportamiento anual y mensual de los incidentes en la ciudad de Medellín.

En una escala temporal más pequeña, la Figura 5-7 muestra que los viernes ($x = 4$) y los sábados ($x = 5$) son los días de la semana con el mayor número de crímenes, mientras que los domingos ($x = 6$) tienen los índices más bajos. Por otro lado, en la escala de horas, se puede observar en la Figura 5-8 que durante días laborales (lunes y martes), existe un gran porcentaje de incidentes para las horas de la mañana y de la tarde correspondientes a los horarios picos de la ciudad. Finalmente, para los días del fin de semana, en la Figura 5-9, se puede evidenciar que hay un gran porcentaje de incidentes entre la noche de los viernes y la madrugada de los sábados, correspondiente a los horarios cuando normalmente las personas se encuentran en bares y discotecas.

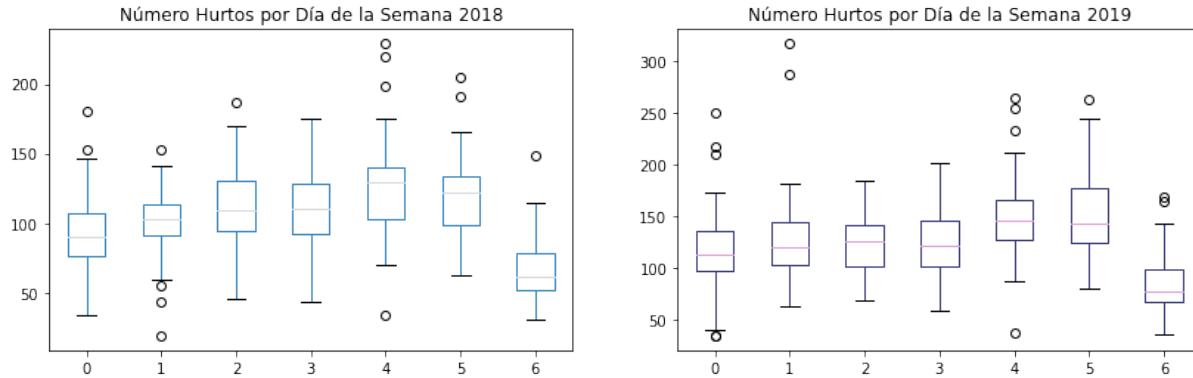


Figura 5-7: Comportamiento semanal de los incidentes en la ciudad de Medellín.

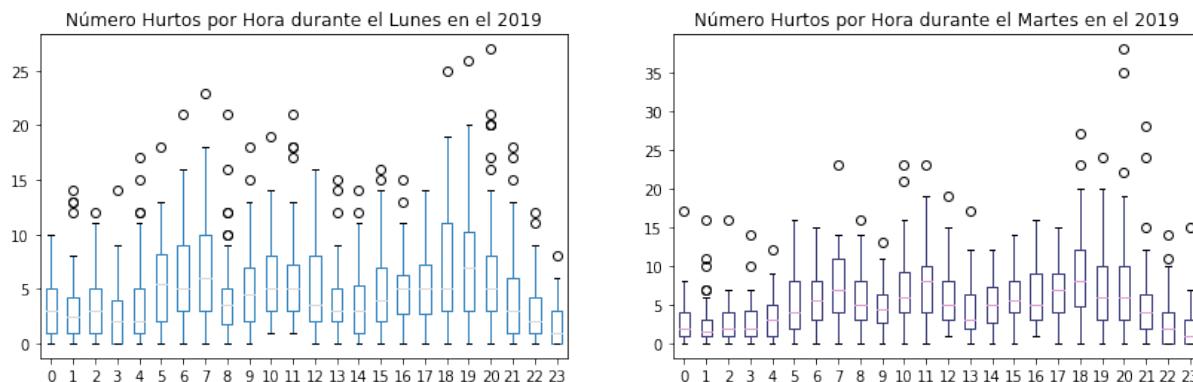


Figura 5-8: Comportamiento por hora de los incidentes en la ciudad de Medellín durante los días laborales.

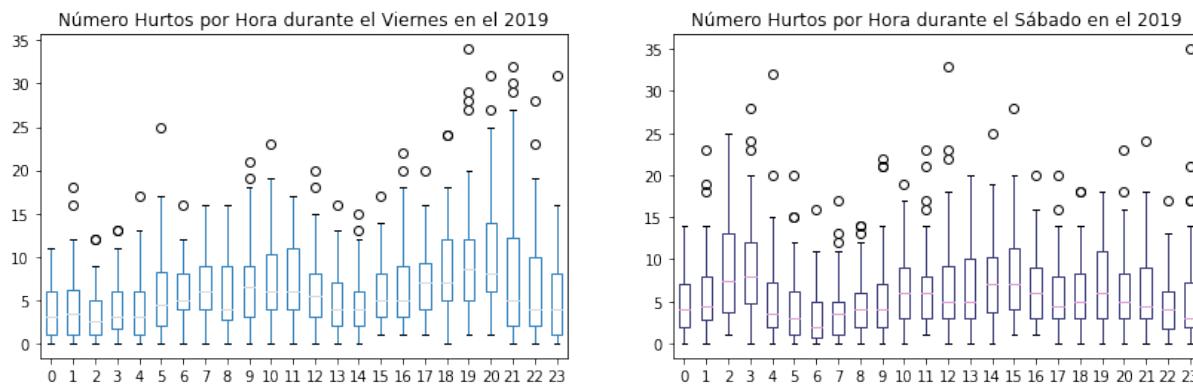


Figura 5-9: Comportamiento por hora de los incidentes en la ciudad de Medellín durante el fin de semana.

■ Información demográfica de la víctima

La base de datos también incluye información demográfica de la víctima. Para iniciar, se encuentra que 57.68 % de los incidentes fueron reportados por hombres, 41.54 % por mujeres y solo el 0.77 % no contienen información. Para el caso de la edad y el estado civil, como se puede observar en la Figura 5-10, la tercera parte de los incidentes que son reportados (33.73 %) corresponden a personas que se encuentran entre los 20 y 29 años, seguidos por un 29.62 % para personas entre los 30 y los 39 años, y un 15.56 % de las personas que se encuentran entre los 40 y 49 años. En este caso, para 1.3 % no se registra un valor para este campo. Por último, en cuanto el estado civil de las personas, es interesante que este campo se almacene en la base de datos y según el análisis, el 55.36 % de los incidentes registrados fueron cometidos sobre personas solteras, un 21.32 % sobre casados y en contraste con los otros resultados se tiene un 4.06 % de los registros sin información. Estas primeras observaciones pueden indicar una falta por parte de las autoridades en el registro del formato, o por su parte, una negativa por parte de las personas a la hora de proveer sus datos personales.

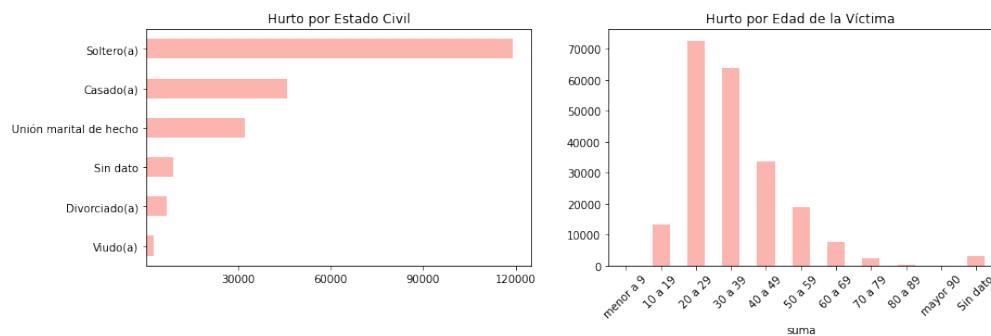


Figura 5-10: Información demográfica de las víctimas.

■ Información del hurto

En cuanto a la información del hurto se puede identificar la modalidad, el medio de transporte y el arma medio. En cuanto a la modalidad cabe resaltar que el 90 % de los incidentes se pueden clasificar dentro de cada una de las siguientes categorías: atraco (49.98 %), descuido (17.62 %), cosquilleo (10.63 %), raponazo (6.83 %), sin dato (4.97 %) y engaño (2.26 %). El resto de los incidentes que se encuentran en la categoría otros (10.01 %) son clasificados por parte de las autoridades en alguna de las siguientes modalidades: Rompimiento cerradura, rompimiento de ventana, escopolamina, halado, clonación de tarjeta, suplantación, comisión de delito, abuso de confianza, informático, llave maestra, paquete chileno, miedo o terror, fleteo, retención de tarjeta, llamada millonaria, forcejeo, simulando necesidad, retención de dinero, tóxico o agente químico, enfrentamiento con la fuerza pública, auto robo, violencia intrafamiliar y vandalismo.

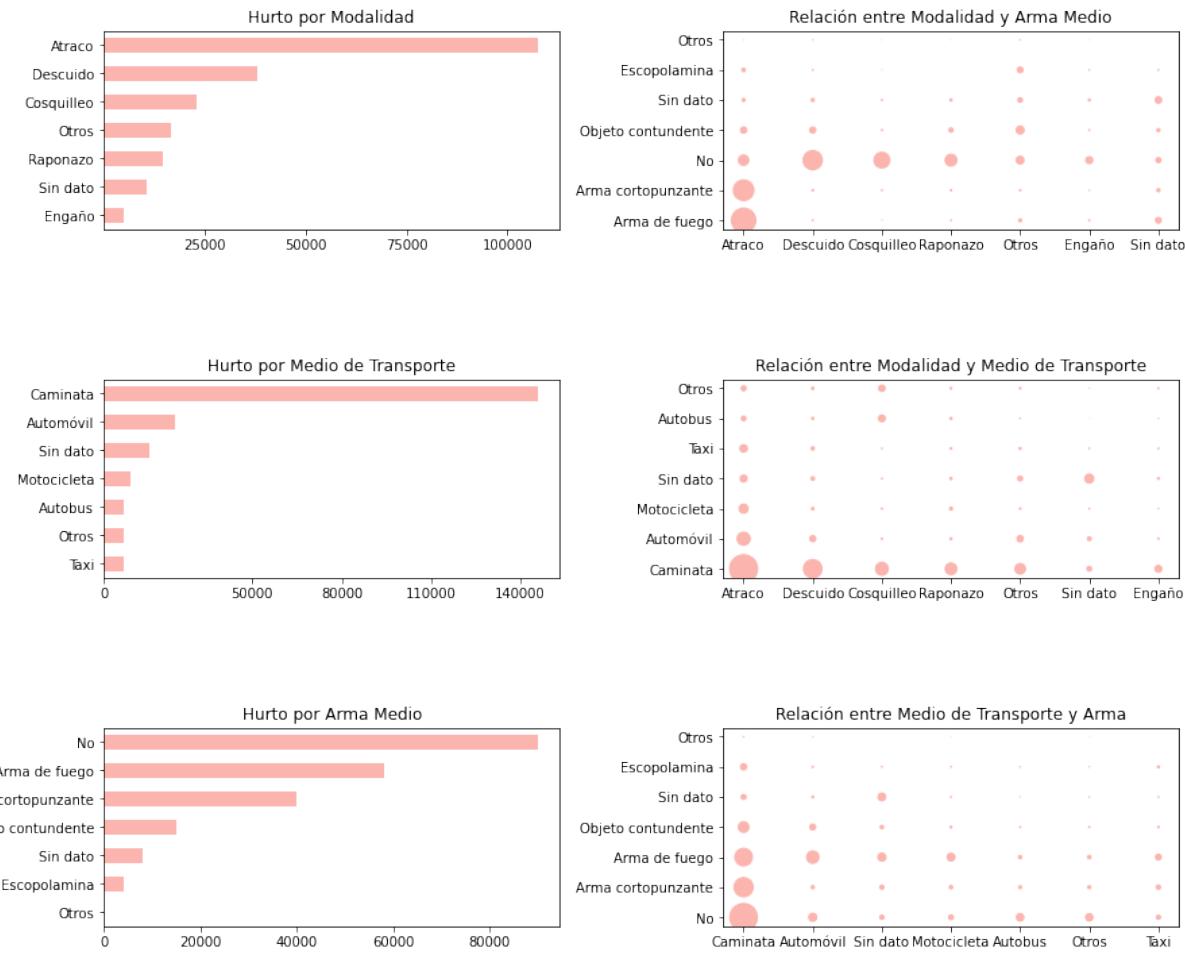


Figura 5-11: Información de las características del hurto en Medellín entre 2003 y 2020.

También se puede encontrar información sobre el medio de transporte en el cual se encontraba la víctima, principalmente se puede notar una gran diferencia con respecto a los incidentes registrados cuando no se estaba en un medio de transporte, es decir cuando la persona se encontraba caminando (67.82 %) con respecto a las personas que se encontraban desplazando en algún medio de transporte. Se puede notar que hay mayor incidencia sobre el transporte privado, como el automóvil (11.19 %) y motocicleta (4.24 %) seguido del transporte público, como el autobús (3.18 %) y el taxi (3.14 %). Por último, hay un porcentaje sin información correspondiente al 7.2 %.

En cuanto el arma medio utilizada por el delincuente, se puede observar que un gran porcentaje de los incidentes se realizaron con la ausencia de un arma medio (41.85 %), y para aquellas en donde el delincuente utiliza un arma, entre las herramientas más utilizadas se encuentra: el arma de fuego (26.96 %), el arma cortopunzante (18.55 %), un objeto contundente (6.95 %) y la escopolamina con un (1.89 %).

Si se realiza un análisis entre la relación de estas tres variables podemos encontrar que el atraco es la modalidad más común, se realiza con arma cortopunzante (17.87 %) y con arma de fuego (25.15 %). De igual forma, se confirma la integridad de la información cuando se reporta que para los delitos de cosquilleo (10.50 %) y descuido de (15.50 %) no se utiliza un arma para su ejecución. Por otro lado, se puede observar que la tercera parte (33.22 %) de los incidentes fueron un atraco en la modalidad de caminata, y aunque el atraco fue la mayor modalidad entre todos los medios de transporte, se confirma una hipótesis popular en donde el cosquilleo se puede ver en mayor porcentaje en el transporte público, este se presenta en mayor porcentaje en el autobús (1.88 %) y en el metro (1.66 %). Para terminar, la mayor parte de los incidentes ocurrieron cuando la víctima se encontraba caminando y no se utilizaron armas (32.13 %), sin embargo, en esta modalidad también se encuentra en gran porcentaje el uso de arma cortopunzante (15.44 %) y arma de fuego (12.94 %). En cuanto la escopolamina, esta técnica fue más utilizada cuando la víctima se encontraba caminando (1.58 %). Finalmente, se observa que el arma de fuego sobresale en número de incidentes cuando la persona se encontraba dentro en de un automóvil (6.37 %) o en una motocicleta (2.59 %).

■ Información del bien robado

En cuanto al bien robado, se construye la Tabla 5-2 con los campos de la base de datos, y en la Figura 5-12 se muestra que el bien más robado es el teléfono celular con un 30.31 % de los incidentes, seguido por peso (Dinero) con un 21.57 %, prendas de

Tabla 5-2: Clasificación de los bienes robados

Grupo Bien	Categoría Bien	Bien
Mercancía	Tecnología	celular, computador, radio, elementos de computador, cámara, tablet
Mercancía	Dinero, joyas, piedras preciosas y título valor	peso (dinero), tarjeta bancaria, dólares
Mercancía	Prendas de vestir y accesorios	accesorios prendas de vestir, billetera zapatos
Mercancía	Documentos	cédula, licencia, libreta militar, soat, tarjeta de identidad, revisión técnico-mecánica
Vehículo	Vehículos de 2 o 4 ruedas	bicicleta, moto, carreta, triciclo
Bélico	Arma de fuego	revólver, pistola, escopeta, pistola neumática, juguete arma bélica
Bélico	Arma blanca	arma blanca, cuchillo, cortopunzantes

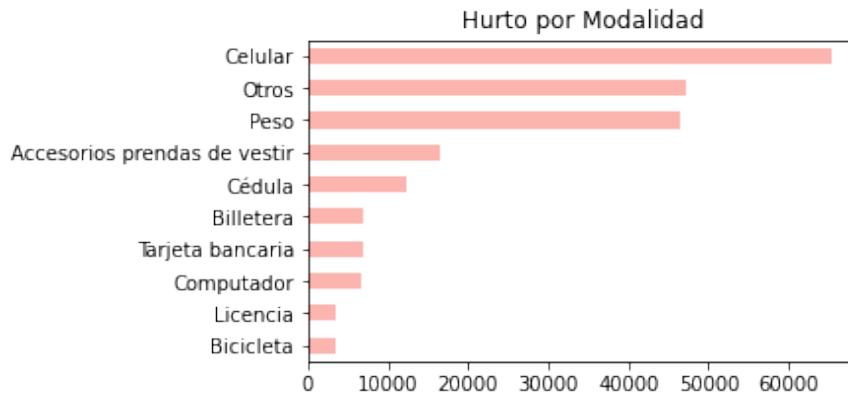


Figura 5-12: Información del bien robado.

vestir con un 7.62 %, y finamente artículos personales como: la cédula 5.72 %, billetera 3.24 % y tarjeta bancaria 3.13 %. Asimismo, se puede observar que para los automóviles la base de datos almacena el modelo (en años) del medio de transporte y el color.

■ Información del lugar

Como se presenta en la Figura 5-13, la mayor concentración de los crímenes se encuentra en el centro de la ciudad, siendo La Candelaria, la comuna con el mayor número de incidentes reportados entre el 2003 y el 2020, aproximadamente una tercera parte (32.36 %). Y en una escala espacial más pequeña, se puede observar que el barrio que recibe el mismo nombre, La Candelaria, abarca el mayor porcentaje de incidentes con un 12.15 %, seguido por los barrios: Colón (2.47 %), el Poblado (2.44 %), Guayaquil (2.22 %) y Villa Nueva (1.92 %).

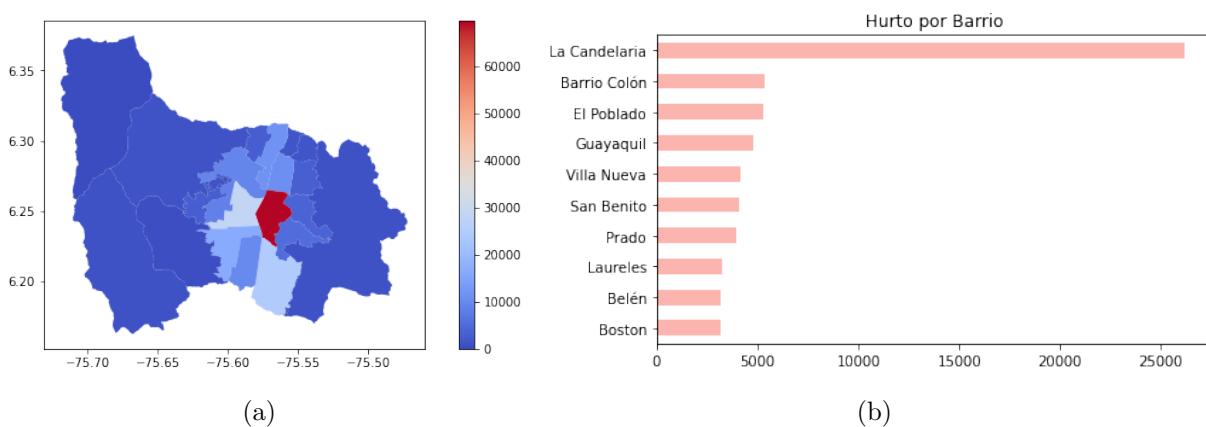


Figura 5-13: Distribución espacial de los crímenes (a) por comuna y (b) por barrio.

Base de datos de las variables demográficas

En este caso, el DANE almacena la Tasa Global de Participación (TGP), la Tasa de Ocupación (TO), la Tasa de Desempleo (TD) y la población total, tanto a nivel nacional, como en las principales ciudades del país, en donde se incluye la ciudad de Medellín. Hay información de enero 2001 hasta diciembre de 2020, y se almacena en forma serie trimestral móvil.

Como se puede observar en la Figura 5-14, la Tasa de Desempleo durante este periodo, posee una media entre el 12 % y 15 %, sin embargo para los meses entre marzo y junio se pueden apreciar unos valores extremos. Tal y como se presenta en la Figura 5-15, estos valores extremos corresponden a la tasa de desempleo en el 2020. En este año, la tasa alcanza un máximo del 25.2 % debido a la Emergencia Sanitaria por COVID-19.

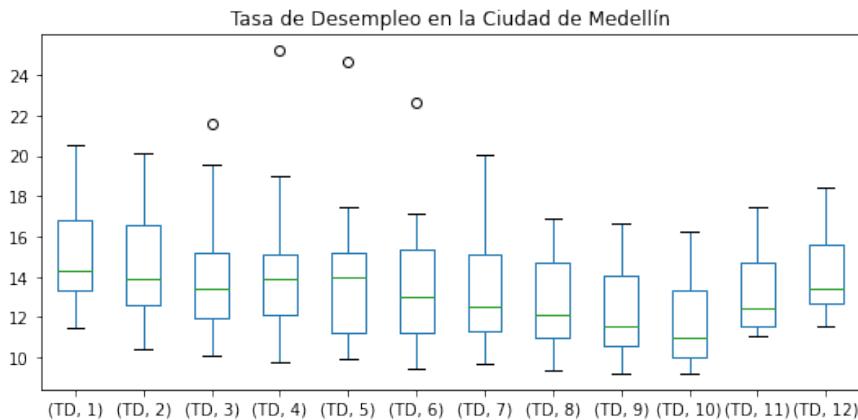


Figura 5-14: Tasa de desempleo en Medellín entre el enero 2000 y diciembre de 2020.

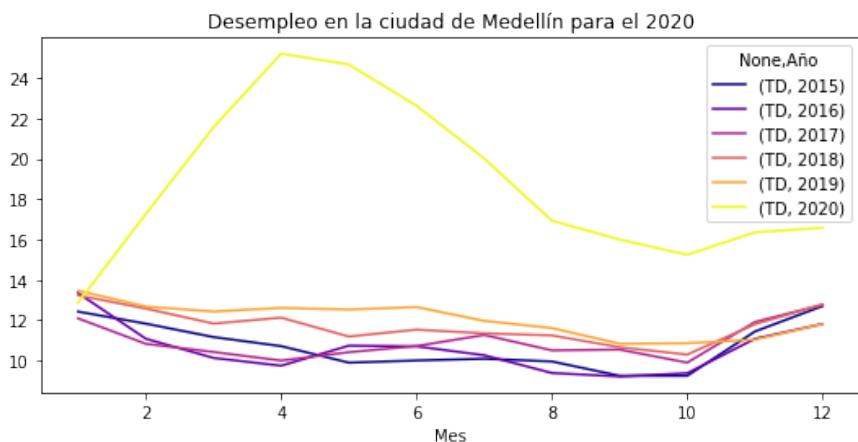


Figura 5-15: Comparación de la tasa de desempleo para en Medellín.

Base de variables meteorológicas

En la revisión de las evidencias se pudo identificar una correlación en algunos países entre los factores metereológico y los incidentes criminales. En el siguiente apartado se pretender revisar y verificar esta relación para la ciudad de Medellín. Para empezar, se realiza un estudio de los recursos dispuestos por el SIATA (Sistema de Alerta Temprana de Medellín y el Valle de Aburrá), quien cuenta con más de 500 estaciones con sensores para variables meteorológicas, hidrológicas y de la calidad del aire repartidas a través de la ciudad. En este momento, la institución almacena información sobre la precipitación, la temperatura, la presión atmosférica, la humedad, la velocidad del viento, y la dirección del viento. En cuanto a las estaciones hidrológicas, monitorea el nivel y la velocidad del río de Medellín. Por último, la red de monitoreo de la calidad del aire se encarga de registrar las concentraciones de los siguientes contaminantes: ozono, monóxido de carbono, monóxido de nitrógeno, dióxido de nitrógeno, óxidos de nitrógeno y dióxido de azufre. Para la presente tesis, nos enfocaremos en las variables meteorológicas, ya que estudios previos indican una posible correlación con la actividad criminal.

Para este estudio, se ha seleccionado una de las estaciones más importantes de la red, la cual contiene sensores para medir: humedad, precipitación, presión atmosférica, velocidad del aire y temperatura. La estación seleccionada es la 202 - AMVA (Área Metropolitana del Valle de Aburrá), la cual se encuentra en la comuna 10 - La Candelaria, el barrio con el mayor número de delitos reportados, y se ha seleccionado una ventana de tiempo entre enero de 2015 y diciembre de 2019. Cada uno de los conjuntos de datos, se compone de una columna de la fecha y hora de registro, el valor de la variable medida y una columna en donde se consigna el índice de calidad. Esta última columna indica que tan confiable es el dato y como se muestra en la Tabla 5-3, este atributo puede tener los siguientes valores [69]:

- 1: Calidad confiable del dato en tiempo real.
- 2: Calidad confiable del dato no obtenido en tiempo real.
- 151: Calidad dudosa en dato de tiempo real.

Tabla 5-3: Calidad de los datos meteorológicos

Calidad	Humedad	Precipitación	Presión	Velocidad del Viento	Temperatura
			Atmosférica		
1	96,73 %	95,40 %	97,00 %	96,73 %	96,73 %
2	0,29 %	0,29 %	0,29 %	0,29 %	0,29 %
151	2,61 %	3,95 %	3,00 %	2,61 %	2,61 %

Tabla 5-4: Estadísticas descriptivas de las variables meteorológicas

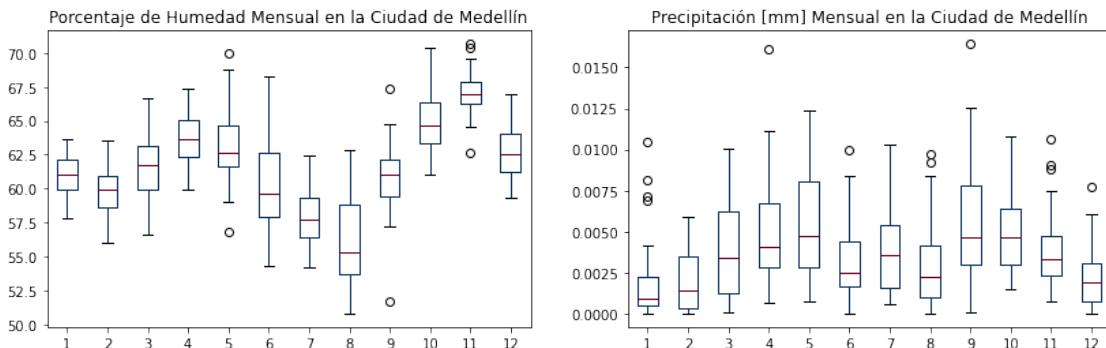
	Humedad	Precipitación	Presión Atmosférica	Velocidad del Viento	Temperatura
Cantidad de muestras	3203780	3248420	3203780	3203780	3203780
Media	61,55 %	0,0037	850	1,66	22,43
Desviación	16,23 %	0,0384	2,43	1,39	3,18
Percentil 50 %	65,60 %	0,0000	850,4	1,3	19,9
Máximo	97,90 %	2,5300	855,9	2,01	32,8

■ Humedad

Como se puede observar en la Tabla 5-4, se obtienen más de 3 millones de registros para el periodo comprendido entre enero de 2015 y diciembre de 2019. Asimismo, se observa que para la ciudad de Medellín, esta variable cuenta con una media de $61,55 \pm 16,23\%$, y el máximo valor que ha alcanzado es del 97,90 %. En cuanto a la calidad de la información, se puede observar en la Tabla 5-3 que el 96,73 % de los datos tienen una calidad confiable del dato en tiempo real. Finalmente, la Figura 5-16, muestra que la humedad alcanza sus máximos relativos en las temporadas entre abril y mayo y entre octubre y diciembre, tal y como se indica en [70].

■ Precipitación

Al igual que la variable anterior, el conjunto de datos para la precipitación se compone de más de 3 millones de datos, con un porcentaje de confiabilidad de los datos del 95,40 %. Para esta variable, se encuentra que en promedio, Medellín tiene precipitaciones de $0,00371 \pm 0,0384 mm$, y su máximo valor ha sido de $2,53 mm$. Al igual que la humedad, se puede observar en la Figura 5-16, que la precipitación también tiene valores máximos durante los periodos entre abril y mayo, y octubre y noviembre.

**Figura 5-16:** Humedad y precipitación en la ciudad de Medellín

■ Presión del aire

Esta variable cuenta con un porcentaje del 97 % de confiabilidad para una muestra de más de 3 millones de registros. Como se indica en la Tabla 5-4, la presión atmosférica en la ciudad de Medellín se encuentra en promedio en $850 \pm 2,43\text{hPa}$, con un valor máximo de $855,9\text{hPa}$. Y en cuanto su comportamiento, en la Figura 5-17 se puede observar un aumento de esta variable entre los meses de abril y septiembre.

■ Velocidad del viento

Para este caso, se cuenta con un conjunto de datos de más de 3 millones de registros y una confiabilidad de los datos del 96,73 %. Como se muestra en la Tabla 5-4, esta variable meteorológica posee un media de $1,66 \pm 1,39\text{m/s}$, y un máximo de $2,01\text{m/s}$. En cuanto a su comportamiento mensual, en la Figura 5-17 se puede observar que existe una gran diferencia entre el mes de agosto con respecto al resto del año.

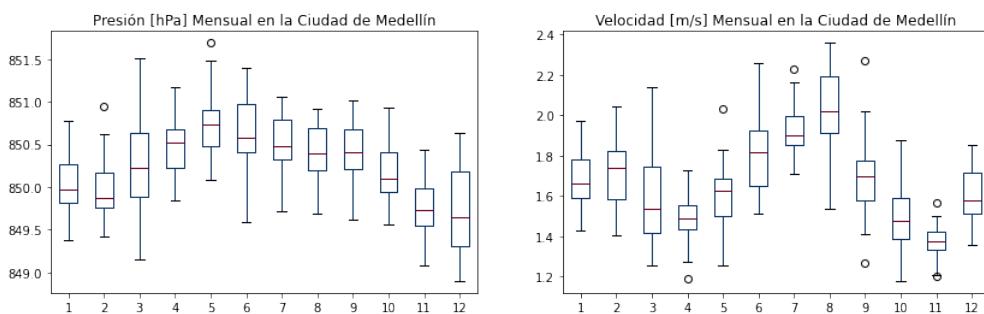


Figura 5-17: Presión Atmosférica y Velocidad del Viento en la ciudad de Medellín

■ Temperatura

Finalmente, para la variable de temperatura, también se cuenta con un conjunto de datos de más de 3 millones de registros y en cuanto a la calidad de la información, se tiene que el 96,73 % de los datos tiene una calidad confiable en el tiempo real. Como se

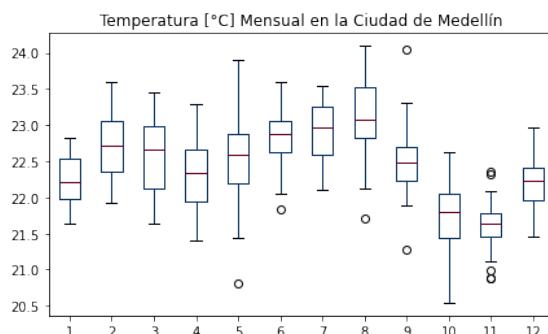


Figura 5-18: Temperatura en la ciudad de Medellín

observa en la Tabla 5-4, en promedio la temperatura en Medellín se encuentra entre $22,43 \pm 3,18^{\circ}\text{C}$. Y como se observa en la Figura 5-18, los meses más fríos del año son octubre y noviembre.

5.3. Preparación de los Datos

5.3.1. Definiciones

Antes de continuar con la construcción del modelo, se procede con la definición de las variables y parámetros que participan en la predicción de zonas calientes [7, 8, 71]. Dado el conjunto $S = \{s_1, s_2, \dots\}$, el cual comprende todas las divisiones espaciotemporales, se pueden determinar las siguientes características para cada división espacial:

$$P^s = \{x_s, y_s\} \quad (5-1)$$

$$M_t^s = \{m_{t-1}^s, m_{t-2}^s, \dots, m_{t-i}^s\} \quad (5-2)$$

$$D_t^s = \{d_{1_{t-1}}^s, m_{d_{t-2}}^s, \dots\} \quad (5-3)$$

Siendo P^s el conjunto de coordenadas, en donde x representa la coordenada de longitud y y la coordenada de latitud. Por su parte, M_t^s representa el histórico del número de delitos m para cada ventana temporal hasta un cierto número de retrasos i , y D_t^s corresponde al histórico de las variables independientes d para cada división. Por otro lado, para cada ventana temporal $S_t \subset S$, se determina el conjunto de las zonas calientes $C_t = \{c_t^{s_1}, c_t^{s_2}, \dots\}$ de la siguiente forma:

$$c_t^s = \begin{cases} 1 & \text{if } m_t^s > h \\ 0 & \text{if } m_t^s \leq h \end{cases} \quad (5-4)$$

$$h = \{5 * \frac{\sum M_t}{N}, \quad \frac{\sum_1^4 m_{t-i}^s}{4}, \quad a, \quad a_t\} \quad (5-5)$$

En donde h representa el umbral de zona caliente y esta definido por alguna de las opciones que se presentan en la Ecuación 5-5. Finalmente, la predicción del crimen se puede conseguir por medio de la solución de una tarea de clasificación binaria dada por la siguiente función:

$$c_t^s = f(M_t^s, D_t^s) \quad (5-6)$$

5.3.2. Selección de los datos

Según la descripción de los datos obtenidos se decide tener en cuenta las siguientes consideraciones:

1. Para la construcción del conjunto de datos final es necesario agrupar los incidentes con las mismas características en el espacio temporal, sin embargo, a medida que se agregan nuevas características a la separación se reduce la cantidad de incidentes disponibles. Por esta razón, se decide utilizar únicamente los atributos relacionados a la fecha y a las coordenadas geográficas, y disponer de los campos con información sobre las características demográficas de la víctima, la información del hurto y las propiedades del bien hurtado.
2. Se decide trabajar con un periodo de tiempo entre enero de 2015 y diciembre de 2019, un periodo de 60 meses, dado que además de ser un periodo de intercesión entre las diferentes fuentes de información, en este periodo es donde se encuentra la mayor concentración de incidentes. Se excluye el año 2020, considerado un año atípico, dada la emergencia sanitaria por COVID-19.
3. Según las distribuciones observadas en la exploración de la base de datos de hurtos, se decide trabajar con aquellas modalidades con los mayores porcentajes, las cuales son: atraco, descuido, cosquilleo y raponazo.
4. Con respecto a las variables independientes, solo se trabajará con aquellas que tengan una afinación con respecto al tamaño de la ventana temporal seleccionada.

5.3.3. Preprocesamiento de los datos

Para la base de datos de hurto a persona se ejecutan los siguientes pasos con la finalidad de tener una buena calidad de la información:

1. Se excluyen todos los registros que no contienen información de latitud o longitud, o que la información que se encuentra en la base de datos no corresponde a las coordenadas que comprende la ciudad de Medellín.
2. Tal como se indica en el apartado sobre la selección de los datos, se realiza un filtrado de las modalidades de hurto: atraco, descuido, cosquilleo y raponazo.
3. Se selecciona el intervalo desde enero de 2015 hasta diciembre de 2019.
4. Se toman la longitud, latitud y la fecha del incidente.

Una vez aplicados todos estos procesos, se obtiene un conjunto de datos de 125.886 registros.

5.3.4. Construcción e integración de información

A partir de las actividades definidas en la comprensión del tema de interés, se ve la necesidad de generar una división espacial para la asignación de todos los crímenes. Para el caso de Medellín, esta separación se puede realizar utilizando las divisiones administrativas de la ciudad como son la comuna y el barrio, o se puede realizar una división por una malla espacial como se muestra en la Figura 5-19.

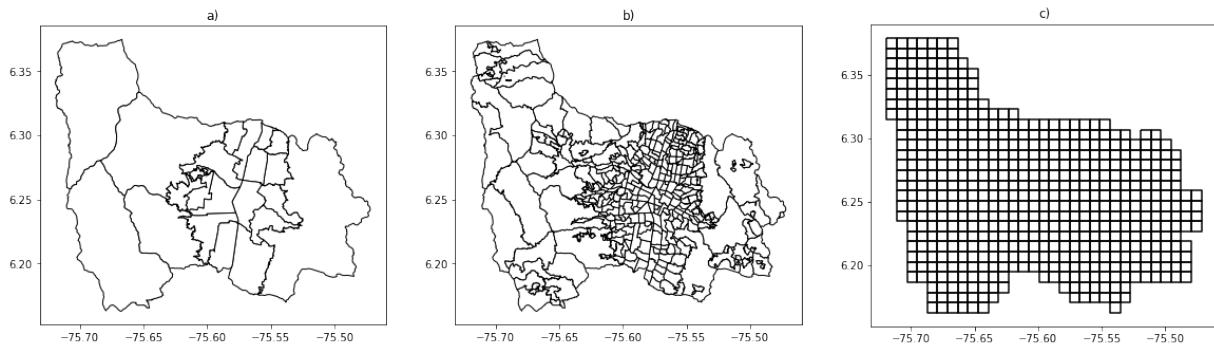


Figura 5-19: Escala espacial disponible para la ciudad de Medellín

Para la presente tesis se decide trabajar con la última división, ya que la malla espacial permite aprovechar las eficiencias en la programación de las operaciones matriciales. En este caso, se utiliza como referencia la superficie promedio de los barrios de Medellín que corresponde a $0,3727\text{km}^2$, resultando en celdas con un lado de $610,5\text{m}$, que equivalen a $0,0055^\circ$. Esta área corresponde al $0,0975\%$ de la superficie total de la ciudad, y con esta división se obtiene una malla espacial de 1775 celdas distribuidas en 45 columnas y 39 filas, pero luego de seleccionar las celdas que se encuentran dentro del área de la ciudad, resulta un total de 1118. A continuación se presenta el código para la generación de la malla espacial, en donde X_{min} , X_{max} , Y_{min} y Y_{max} corresponden a los límites perimetrales de la ciudad y *polygons* es el arreglo en donde se almacenan las coordenadas de cada celda:

Algoritmo I

```

height = 0,0055
width = 0,0055
polygons = []
cols = int(np.ceil((Xmax - Xmin)/width))
rows = int(np.ceil((Ymax - Ymin)/height))
Xleft,origin = Xmin
Xright,origin = Xmin + width
Ybottom,origin = Ymin
Ytop,origin = Ymin + height

```

```

for i in range(rows):
    xleft = Xleft,origin
    xright = Xrigth,origin
    for j in range(cols) :
        polygons.append(Polygon([(Xleft, Ytop,origin), (Xright, Ytop,origin),
                                (Xright, Ybottom,origin), (Xleft, Ybottom,origin)]))
        Xleft = Xleft + width
        Xright = Xright + width
        Ybottom,origin = Ybottom,origin + height
        Ytop,origin = Ytop,origin + height

```

Como paso siguiente, se realiza la enumeración de todas las celdas de izquierda a derecha y de abajo hacia arriba, y se procede con la asignación de cada uno de los incidentes a cada una de las celdas que hacen parte de la malla espacial. Gracias a la distribución matricial, se utiliza la siguiente ecuación para determinar el número de celda al que pertenece cada incidente:

$$\#celda = \text{int} \left(\frac{\text{longitude} - Y_{\min}}{\text{height}} \right) * \text{cols} + \text{int} \left(\frac{\text{latitude} - X_{\min}}{\text{width}} \right) \quad (5-7)$$

Mediante la programación se procede de la siguiente manera, Siendo dfCoord el Dataframe con todas las longitudes y latitudes almacenadas en el campo *geometry*:

Algoritmo II

```

dfCoord["x"] = dfCoord["geometry"].apply(lambda x : x.latitude)
dfCoord["y"] = dfCoord["geometry"].apply(lambda x : x.longitude)
dfCoord["xo"] = dfCoord["x"].apply(lambda x : int((x - xmin)/width))
dfCoord["yo"] = dfCoord["y"].apply(lambda x : int((x - ymin)/height))
dfCoord["Cell"] = dfCoord["yo"].apply(lambda x : x * cols) + dfCoord["xo"]

```

Una vez asignados todos los incidentes, se puede observar en la Figura 5-20 que solo 523 (46 %) de las celdas contienen valores para el período comprendido entre enero de 2015 y diciembre de 2019. Estas celdas cuentan con una media de 240 hurtos y una desviación que supera el doble de este valor con un total de 705 incidentes. Esto último se puede explicar por la gran diferencia que se pudo apreciar en el apartado de la exploración, cuando se estableció que el centro de la ciudad concentra alrededor de la tercera parte del total de todos los incidentes.

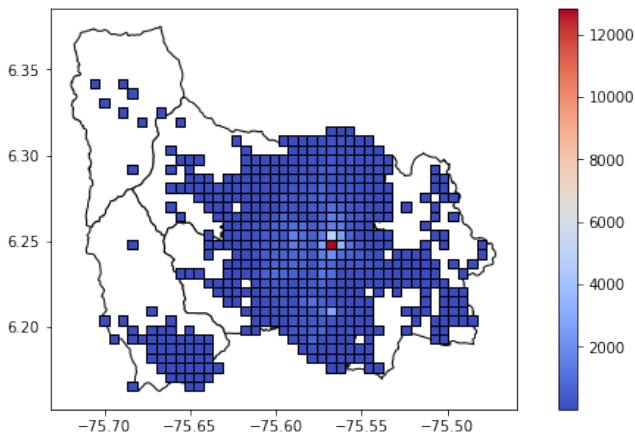


Figura 5-20: Distribución de hurto en la ciudad de Medellín entre el 2015 y 2019

Tabla 5-5: Estadísticas descriptivas de la distribución de celdas

mean	std	25 %	50 %	75 %	95 %	max
240	705	4	44	263	919	12842

En la Tabla 5-5 se puede apreciar la distribución de las estadísticas descriptivas de la muestra, y es claro que solo se deberían considerar celdas con un número considerable de incidentes, es decir, aquellas que en algún momento se puedan convertir en zonas calientes, y que tampoco sean valores extremos. A forma de comparación, si solo se consideran las celdas con más de 50 y menos de 1000 incidentes (entre los percentiles 50 y 95) se puede observar en la Figura 5-21 una mejor distribución de los hurtos, con una media de 225 y una desviación estándar de 310.

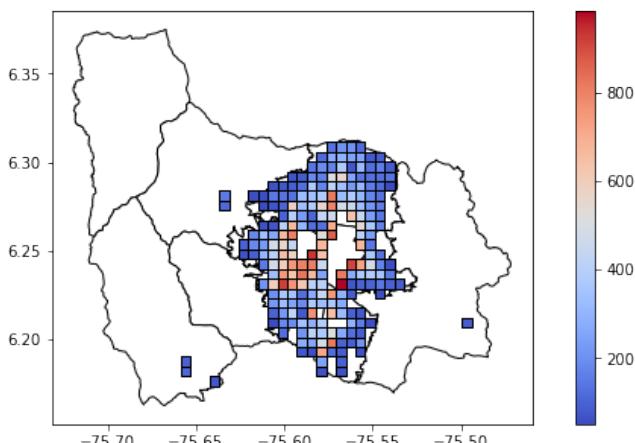


Figura 5-21: Distribución de hurto en la ciudad de Medellín entre el 2015 y 2019

Dadas estas divisiones espaciales, el siguiente paso consiste en realizar una división temporal que permitirá la construcción de las características para el conjunto de datos final. Como se muestra en la Tabla Tabla 5-6 la selección de este parámetro representa un compromiso entre el número de muestras disponibles para el entrenamiento y la calidad de la información que se encuentra al interior de cada ventana espaciotemporal. Para valores muy pequeños como 1 día o una semana, se observa una gran cantidad de celdas para la predicción de crímenes, sin embargo, por cada división existen valores muy pequeños, entre 0 o 1 incidente. Para el caso de ventanas más grandes como 4 semanas se puede observar que la calidad de la información mejora, con una media de 4 crímenes por división espaciotemporal y con un percentil 95 de 16, sin embargo, comparando con respecto a las divisiones más pequeñas se puede observar una reducción considerable del número de muestras disponibles para el entrenamiento.

Tabla 5-6: Comparación de diferentes valores para la ventana temporal

Variable	1 Día	1 Semana	2 Semanas	3 Semanas	4 Semanas
Ventanas Temporales	1825	262	132	88	67
Total de Divisiones	954475	137026	69036	46024	35041
Promedio de Incidentes	0	1	2	3	4
Percentil 95 %	1	5	9	13	16
Máximo de Incidentes	93	135	241	321	449
Celdas con Incidentes por Ventana Temporal	32	125	174	204	224
% Celdas con Incidentes por Ventana Temporal	6,15 %	23,95 %	33,32 %	39,06 %	42,38 %

La selección de este parámetro debería ser consecuente con las estrategias de las autoridades, en este punto, se puede relacionar la división temporal con los lineamientos de rotación de los recursos en la ciudad. Para este caso de estudio, se decide utilizar un valor óptimo con fines teóricos, una ventana temporal de 2 semanas, ya que tiene un buen porcentaje de muestras para el entrenamiento de los modelos, comparada con la escala de 3 semanas y 4 semanas, y que aunque el promedio de incidentes por cada división es de 2, el percentil 95 está definido en 9, suponiendo que las zonas calientes se determinan a partir de este percentil, cada quincena se podrían estar gestionando al menos 9 incidentes por cada grupo de patrullaje.

Con esta división se obtiene un total de 132 ventanas temporales y con las 523 celdas, la base de datos resultante contiene un total de 69036 registros para el entrenamiento de los modelos. El promedio de incidentes en todas las celdas es de aproximadamente 2 incidentes con una desviación de 6 crímenes, siendo 241 el número máximo de crímenes reportados en una división espaciotemporal. Finalmente, para cada una de las ventanas temporales, el promedio de celdas que contiene delitos es de 174 celdas con una desviación de 37 celdas.

El siguiente paso consiste en la selección de una directriz para la clasificación de las divisiones que corresponden a una zona caliente para cada ventana temporal. Como se mostró en la sección de las definiciones, para determinar una zona caliente se debe definir un umbral que puede utilizar información como el promedio de delitos dentro de una zona temporal o el número de crímenes históricos. Con el objetivo de incorporar esta definición con una posible estrategia de las autoridades, se decide establecer un nuevo parámetro K que representa el número de recursos disponibles por parte de las autoridades. De modo que la definición de zona caliente para la presente tesis queda representado por la siguiente ecuación:

$$h = \text{TOP}_k(M_k) \quad (5-8)$$

Así como se muestra en la figura 5-22, la selección de este nuevo parámetro representa un compromiso entre la calidad de las zonas calientes a predecir, y el porcentaje del área cubierta en la ciudad de Medellín. Para el caso de $K = 10$ se observa que solo se está cubriendo un 20% de todas las zonas de la ciudad, mientras con un $K = 30$, se obtiene una cobertura del 43% de la ciudad. Para este caso se decide utilizar un paso del 20% dado que se mantiene un calidad adecuada de zonas calientes, y en cuanto al porcentaje de cobertura se puede observar que entre $K = 10$ y $K = 20$ hay un aumento del 15% mientras que la diferencia entre este último y $K = 30$ solo es de 8%.

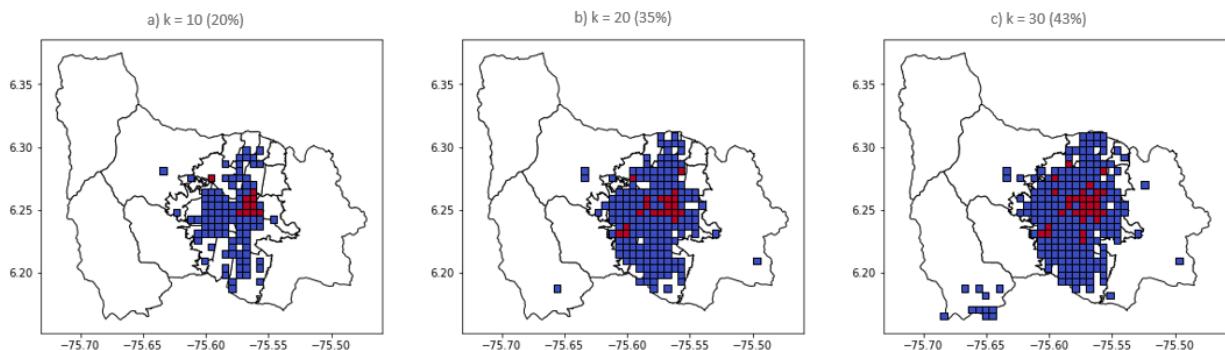


Figura 5-22: Comparación para diferentes valores del umbral y área cubierta de la ciudad.

Se procede entonces con el etiquetado de los registros que corresponden a zonas calientes. Como resultado, se obtiene un total de 183 celdas que corresponden a una zona caliente para cualquiera de las ventanas temporales. En la Figura 5-23 se puede observar la distribución del número de veces que cada celda es clasificada como una zona caliente. Para la presente distribución se obtiene:

- 15 % (28 celdas) son clasificadas como zonas calientes al menos para 50 períodos
- 48 % (88 celdas) son clasificadas como zonas calientes al menos para 10 ventanas temporales.

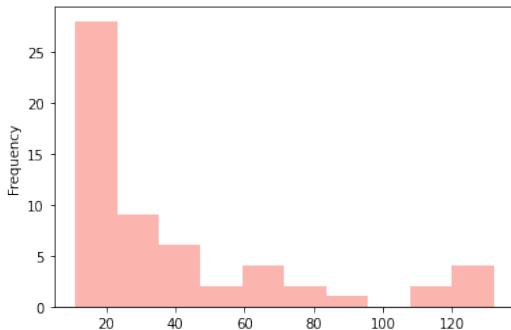


Figura 5-23: Distribución de zonas calientes por celda

Dado que las divisiones espaciales que no corresponden a una zona caliente no proporcionan información relevante al modelo, se decide disponer de estas celdas, reduciendo las muestras a 24156 divisiones espaciotemporales. Este último porcentaje corresponde a la cobertura que se muestra en la figura 5-22.

Hasta este punto se tiene un conjunto de datos con los siguientes campos: fecha de la ventana temporal, número de celda, número de hurtos y su respectiva etiqueta indicando si corresponde a una zona caliente. El siguiente paso consiste en generar las características para el entrenamiento del modelo. Para iniciar, se decide crear una serie de retrasos del comportamiento de hurtos para cada una de las muestras, es decir, para cada registro se calcula el comportamiento de los hurtos en las ventanas anteriores. En este caso, se decide hacer un proceso repetitivo de los 10 retrasos contiguos a la zona caliente que se está estudiando. Por último, con respecto a las variables independientes, se decide trabajar solo con la variable de desempleo. Lo anterior, por los resultados reportados en [72], en donde se evidencia que para la ciudad de Medellín existe una relación directa entre el desempleo y el número de crímenes, mientras que para las variables metereológicas no se encuentra una relación directa.

El conjunto de datos final se compone de 22326 registros y 12 campos que corresponden a cada uno de los retrasos al evento que se quiere predecir, la tasa de desempleo y una etiqueta que indica si la división espacio temporal corresponde a una zona caliente. Ahora bien, para el problema de clasificación binaria planteado existe un desafío para el entrenamiento de los modelos, dado por la definición del umbral que genera un desequilibrio entre las zonas calientes y las zonas frías [6]. Específicamente para esta investigación, solamente el 10.92 % de los registros corresponden a zonas calientes. Por consiguiente, con el propósito de mejorar el entrenamiento de los modelos y evitar el sesgo de los mismos hacia las zonas frías [73], se decide realizar un submuestreo, una selección de datos aleatorios de la etiqueta que no corresponde a una zona caliente. Con esto, el total de muestras se reduce a 4880.

6 Modelos para la Predicción del Crimen en la Ciudad de Medellín

Con el objetivo de validar la metodología planteada en la presente tesis, se propone realizar una validación de cuatro modelos, uno basado en reglas y tres de *Machine Learning*, para la predicción de hurtos en la ciudad de Medellín. Por lo tanto, en este capítulo se procederá con la enumeración de los modelos, la explicación sobre la técnica de evaluación, la construcción de los modelos y la presentación de los resultados.

6.1. Selección de la Técnica de Modelado

Basados en la metodología planteada, para la presente tesis se realiza una comparación entre cuatro modelos, un modelo base basado en reglas, y tres modelos de *Machine Learning*. El primer modelo basado en reglas, sirve para comparar un acercamiento para tratar este mismo tipo de problemas, pero sin utilizar las herramientas que ofrece el aprendizaje automático. Para esto, se desarrollan unas reglas basadas en el conjunto de datos construido, con el fin de determinar si una división espaciotemporal corresponde a una zona caliente. Este primer modelo servirá como punto de partida para la comparación de desempeño de los modelos restantes. Por otro lado, en cuanto a los modelos de aprendizaje automático, se deben seleccionar modelos para un problema supervisado (conjunto de datos con etiquetas) y para una inferencia que corresponde a una tarea de clasificación (salida discreta). En este trabajo se presentan los modelos más utilizados en la literatura para este tipo de tareas:

- Modelo Base - Basado en reglas
- Regresión Logística (*Logistic Regression*)
- Bosques Aleatorios (*Random Forest*)
- Máquina de Soporte Vectorial (*Support Vector Machine*)

6.2. Técnica de Evaluación

En cuanto a las técnicas de evaluación se contemplan dos procesos:

- Para el método basado en reglas, dado que no es necesario realizar una calibración de parámetros, se aplican las reglas a todo el conjunto de datos y se procede con el cálculo de las métricas de desempeño más utilizadas para los modelos de clasificación: exactitud, exhaustividad, precisión y el valor F1.
- Para el caso de los algoritmos de aprendizaje automático, se utiliza la técnica de Validación Cruzada Doble (*Nested Cross Validation*), con el fin de realizar la calibración de los hiperparámetros, al mismo tiempo que se valida el desempeño de los modelos. Para el ciclo externo en donde se lleva a cabo la validación de los mejores modelos, se realiza una partición en 10 partes del conjunto de datos, mientras que para el ciclo interno, en donde se efectúa la calibración de los parámetros, se realiza una partición de 3 unidades. Como métrica para la calibración de los parámetros para la selección del mejor modelo, se utiliza una función de costo para la métrica de *valorF1* y para la validación de los modelos se calculan todas las métricas en el flujo externo, para luego compararlas con el modelo base.

6.3. Construcción del Modelo

A continuación se listan todos los parámetros que son utilizados para la calibración y construcción de cada uno de los modelos:

- Modelo basado en reglas: como se tiene establecido, el conjunto de datos contiene la información del número de hurtos de los 10 períodos anteriores a la fecha del incidente, la tasa de desempleo y la etiqueta indicando si la división espacio temporal corresponde a una zona caliente. Como modelo base, se establecen las siguientes reglas para la predicción de zonas calientes:
 - i. Si el promedio de incidentes en las últimas 10 ventanas de tiempo supera el percentil 95, se clasifica como zona caliente.
 - ii. Si el promedio de los incidentes en las últimas 5 ventanas temporales supera el percentil 95, se clasifica como zona caliente.
 - iii. Si todas las ventanas temporales contienen información sobre hurtos, se clasifica como zona caliente.
 - iv. En caso contrario, se clasifica como ausencia de zona caliente
- Regresión Logística [74]:
 - Algoritmo para resolver el problema de optimización: newton-cg, lbfgs, liblinear
 - El inverso de la constante de Regularización C : 100, 10, 1, 0.1, 0.01
 - Penalidad: l2

- Bosques Aleatorios [75]:
 - Número árboles o estimadores en el bosque ($n_{estimators}$): 10, 100, 1000
 - Número de características en cada partición ($max_{features}$): 'sqrt', 'log2'
- Máquina de soporte vectorial [76]:
 - Función Kernel: poly, rbf, sigmoid
 - Constante de Regularización C : 50, 10, 1, 0.1, 0.01

6.4. Resultados

En esta sección se presentan los resultados de la validación cruzada doble aplicada a los modelos de aprendizaje automático descritos en el presente capítulo, así como la comparación de su desempeño contra el modelo basado en reglas. Es de recordar que las métricas que se utilizaron para evaluar los modelos son la exactitud, la precisión, la exhaustividad y el valor F1. Los resultados se reportan de la siguiente manera: $promedio \pm desviacion\%$.

- Regresión logística

Como se puede observar en la Tabla 6-1, con el modelo propuesto de la regresión logística se logra obtener una exactitud de $75,53 \pm 1,53\%$, una exhaustividad de $66,27 \pm 2,39\%$, un valor F1 de $73,01 \pm 1,88\%$ y una precisión de $81,38 \pm 2,69\%$. Y en cuanto a los mejores hiperparámetros destacan una constante C de 100 y el algoritmo de solución *newton-cg*.

Tabla 6-1: Resultados de la Validación Cruzada Doble sobre Regresión Logística

i	Exactitud	Exhaustividad	Valor F1	Precisión	Mejor C	Mejor Solver
1	0.7254	0.6540	0.6982	0.7488	100	'newton-cg'
2	0.7439	0.6590	0.7335	0.8269	1	'newton-cg'
3	0.7520	0.6224	0.7126	0.8333	1	'liblinear'
4	0.7541	0.6573	0.7309	0.8232	100	'newton-cg'
5	0.7561	0.6723	0.7289	0.7960	100	'newton-cg'
6	0.7643	0.6627	0.7416	0.8418	1	'newton-cg'
7	0.7766	0.7039	0.7506	0.8039	100	'newton-cg'
8	0.7643	0.6880	0.7495	0.8230	100	'newton-cg'
9	0.7398	0.6276	0.7026	0.7979	1	'newton-cg'
10	0.7766	0.6803	0.7528	0.8426	100	'newton-cg'
Promedio	0.7553	0.6627	0.7301	0.8138		
Desviación	0.0153	0.0239	0.0188	0.0269		

Bosques Aleatorios

Para el caso de los bosques aleatorios, en la Tabla **6-2** se presentan los resultados obtenidos. En este caso, el modelo obtiene una exactitud de $76,25 \pm 1,30\%$, una exhaustividad de $71,87 \pm 2,30\%$, un valor F1 de $75,13 \pm 1,74\%$ y una Precisión de $78,75 \pm 2,04\%$. Y en cuanto a los mejores hiperparámetros se pueden encontrar con mayor frecuencia el uso de 1000 árboles y el método de *sqrt* para la separación de los atributos.

Tabla 6-2: Resultados de la validación cruzada doble sobre bosques aleatorios

i	Exactitud	Exhaustividad	Valor F1	Precisión	Mejor nEstimators	Mejor maxFeatures
1	0.7480	0.7046	0.7309	0.7591	1000	'sqrt'
2	0.7561	0.7241	0.7606	0.8008	1000	'sqrt'
3	0.7664	0.6929	0.7455	0.8068	1000	'sqrt'
4	0.7623	0.7056	0.7511	0.8028	100	'sqrt'
5	0.7643	0.7437	0.7548	0.7662	1000	'sqrt'
6	0.7807	0.7510	0.7775	0.8060	1000	'sqrt'
7	0.7705	0.7382	0.7544	0.7713	1000	'sqrt'
8	0.7684	0.7280	0.7631	0.8018	1000	'sqrt'
9	0.7336	0.6736	0.7124	0.7559	1000	'sqrt'
10	0.7746	0.7254	0.7629	0.8045	1000	'sqrt'
Promedio	0.7625	0.7187	0.7513	0.7875		
Desviación	0.0130	0.0230	0.0174	0.0204		

Máquinas de Soporte Vectorial

Finalmente, para el caso de la máquina de soporte vectorial, en la Tabla **6-3** se presentan los resultados obtenidos. En este caso, el modelo obtiene una exactitud de $74,67 \pm 1,98\%$, una exhaustividad de $80,57 \pm 2,69\%$, un valor F1 de $76,06 \pm 2,03\%$ y una precisión de $72,11 \pm 2,78\%$. En cuanto a los mejores hiperparámetros se pueden encontrar con mayor frecuencia, un parámetro de regularización *C* de 50 y un kernel en función de polinomio.

Análisis

Como se puede observar en la tabla **6-4** el modelo con la mejor métrica de desempeño de valor F1 corresponde al modelo de máquina de soporte vectorial, con un puntaje del 76,06 %, seguido por el algoritmo de bosques aleatorios con 75.13 % y la regresión logística con un total de 73.01 %. En términos de la exhaustividad, el modelo con el mayor puntaje fue obtenido por bosques aleatorios 76.25 %, seguido por el modelo de regresión logística 75,53 % y por

Tabla 6-3: Resultados de la validación cruzada doble sobre SVC

i	Exactitud	Exhaustividad	Valor F1	Precisión	Mejor C	Mejor Kernel
1	0.7111	0.7932	0.7273	0.6714	1	'poly'
2	0.7439	0.8046	0.7706	0.7394	50	'poly'
3	0.7623	0.7759	0.7633	0.7510	50	'poly'
4	0.7398	0.7823	0.7534	0.7266	50	'poly'
5	0.7459	0.8613	0.7678	0.6926	1	'poly'
6	0.7664	0.8273	0.7833	0.7437	50	'poly'
7	0.7459	0.7983	0.7500	0.7072	50	'poly'
8	0.7643	0.8120	0.7793	0.7491	10	'poly'
9	0.7152	0.7699	0.7258	0.6866	50	'poly'
10	0.7725	0.8320	0.7853	0.7436	10	'poly'
Promedio	0.7467	0.8057	0.7606	0.7211		
Desviación	0.0198	0.269	0.0203	0.0278		

el modelo base 74.86 %. Para el caso de la métrica de precisión, tenemos que el modelo de regresión logística tiene el mejor desempeño con un puntaje del 81.38 % seguido por el modelo de bosques aleatorios con 78.75 % y por el modelo base con un puntaje de 75.19 %.

Tabla 6-4: Métricas de los modelos de predicción

Modelo	Exactitud	Exhaustividad	Valor F1	Precisión	Tiempo Entrenamiento
Modelo Base	0.7486	0.7418	0.7468	0.7519	No Aplica
Regresión Logística	0.7553	0.6627	0.7301	0.8138	8 seg
Bosques Aleatorios	0.7625	0.7187	0.7513	0.7875	5 min 20 seg
Máquina de Soporte Vectorial	0.7467	0.8057	0.7606	0.7211	3 min 40 seg

En conclusión, no se cuenta con un modelo que supere a los demás en todas las métricas de desempeño, por lo que la selección del mejor modelo se debe realizar basado en las necesidades y los recursos de las autoridades. Si la necesidad que se tiene consiste en erradicar el máximo número los hurtos alrededor de la ciudad, sin importar el número de recursos utilizados, el modelo de máquina de soporte vectorial sería el más conveniente. Esto debido a que su métrica de exhaustividad es la más alta entre los resultados 80.57 %, indicando que con este modelo se abarcarián más zonas calientes. Si por el contrario, la visión de las autoridades es un poco más austera, y no quieren malgastar los recursos sin importar cuantas zonas calientes dejen de percibir, el modelo de regresión logística es el más indicado, ya que este tiene la precisión más alta con un valor del 81.38 %, sin embargo, esta ventaja tiene su

costo, y esto se ve reflejado en el valor de la exhaustividad en donde se puede apreciar un valor del 66.27 %, el más bajo de todos, incluso, más bajo que el modelo base. Finalmente, si lo que se busca, es una solución balanceada, un modelo que tenga en cuenta el número de zonas calientes que se van a cubrir y la cantidad de recursos disponibles, la máquina de soporte vectorial ofrece la mejor métrica de valor F1 76.06 %, cabe recordar que esta métrica es una combinación entre la exhaustividad y la precisión.

Como se puede observar, el modelo basado en reglas presenta unas métricas apropiadas considerando que no es un modelo de *Machine Learning*, con una exhaustividad de 74.18 % y un Valor F1 de 76.06 % supera algunos modelos. Considerando que estas métricas son altas, se podría pensar en la implementación de un modelo en reglas para la predicción de crímenes en la ciudad, sin embargo, considerando las métricas sobre la percepción de la seguridad en la ciudad de Medellín, al menos el 50 % de las personas se sienten inseguras, y bajo un lineamiento de querer mejorar este indicador, sería adecuado utilizar un modelo de *Machine Learning* como el de maquina de soporte vectorial.

En cuanto al modelo de validación cruzada doble, se puede evidenciar en los resultados obtenidos que no se presenta un sobreajuste de los modelos. Cabe recordar que este método de validación evita el sesgo por los datos utilizados para el entrenamiento, al mismo tiempo que se realiza una comparación de los mejores parámetros en cada uno de los ciclos de validación. Dado que el fin de esta investigación es validar los modelos, no se realiza una comparación con datos externos.

Finalmente, como se pudo observar a lo largo de este investigación, la selección de todos los parámetros se realizó basado en unos supuestos teóricos. Se espera que estos resultados se puedan integrar a las estrategias de la policía en una etapa posterior. Para realizarlo, se deberían seleccionar los siguientes parámetros: la división espacial, la división de la ventana temporal (asociada a la rotación de los recursos), el tipo de delito, las variables independientes y el número de zonas calientes que se desean detectar, esto último con base en la cantidad de recursos disponibles. Todo esto, coordinado con la métrica de predicción que se desea utilizar para las modelos, la cual dependerá de los objetivos de las autoridades, es decir, si desean ser eficientes con la locación de los recursos o si desean detectar la mayor cantidad de zonas calientes.

Reporte Final

Finalmente, en la Tabla 6-5 se reporta un resumen de los parámetros utilizados para el desarrollo de este problema de investigación mediante la metodología propuesta.

Tabla 6-5: Resumen de la metodología aplicada

ID	Fase	Tarea	Resultado	Estudio de Caso
1	Comprensión del Tema de Interés	Determinar los Objetivos del Tema de Interés	<ul style="list-style-type: none"> • Patrocinador/Beneficiarios • Objetivos • Área de Estudio (Escala)* 	<ul style="list-style-type: none"> • Autoridades de Medellín • ver Capítulo 5.1 • Ciudad de Medellín
		Evaluación de la Situación	<ul style="list-style-type: none"> • Inventario de Recursos 	<ul style="list-style-type: none"> • ver Capítulo 5.1
		Determinar los Objetivos de Minería de Datos	<ul style="list-style-type: none"> • Objetivos de Aprendizaje Automático • Tipo de Inferencia* 	<ul style="list-style-type: none"> • ver Capítulo 5.1 • Predicción de Zonas Calientes
		Generar un Plan de Proyecto	<ul style="list-style-type: none"> • Lista de Actividades 	<ul style="list-style-type: none"> • ver Capítulo 5.1
2	Comprensión de los Datos	Recopilación de los Datos	<ul style="list-style-type: none"> • Reporte de la Recopilación 	<ul style="list-style-type: none"> • ver Capítulo 5.2.1
		Descripción de los Datos	<ul style="list-style-type: none"> • Reporte de la Descripción 	<ul style="list-style-type: none"> • ver Capítulo 5.2.2
		Exploración de los Datos	<ul style="list-style-type: none"> • Reporte de la Exploración 	<ul style="list-style-type: none"> • ver Capítulo 5.2.2
		Verificación de la Calidad	<ul style="list-style-type: none"> • Reporte de la Calidad de Datos 	<ul style="list-style-type: none"> • ver Capítulo 5.2.2
3	Preparación de los Datos	Selección de los Datos	<ul style="list-style-type: none"> • Motivos de Inclusión y Exclusión • Periodo de Tiempo (Meses)* • Tipo de Crimen* • Variables 	<ul style="list-style-type: none"> • ver Capítulo 5.3.1 • Ene. 2015 a Dic. 2019 (60 meses) • Atraco, Descuido, Cosquilleo, Raponazo • Desempleo
		Preprocesamiento	<ul style="list-style-type: none"> • Reporte del Preprocesamiento 	<ul style="list-style-type: none"> • ver Capítulo 5.3.2
		Construcción de los Datos	<ul style="list-style-type: none"> • Lista de Atributos Generados • Lista de Registros Generados 	<ul style="list-style-type: none"> • ver Capítulo 5.3.3 • ver Capítulo 5.3.3
		Integración de los Datos	<ul style="list-style-type: none"> • Lista de Integraciones Realizadas 	<ul style="list-style-type: none"> • ver Capítulo 5.3.3
		Formateo de los Datos	<ul style="list-style-type: none"> • Formato de los Datos 	<ul style="list-style-type: none"> • ver Capítulo 5.3.3
			<ul style="list-style-type: none"> • Conjunto de Datos Final • Tamaño de la Muestra* • Escala Espacial* • Escala Temporal* 	<ul style="list-style-type: none"> • ver Capítulo 5.3.3 • 4880 • 610 m x 610 m • 2 Semanas
4	Modelado	Selección de la Técnica de Modelado	<ul style="list-style-type: none"> • Tarea de Predicción* • Modelo Base* • Lista de Modelos* 	<ul style="list-style-type: none"> • Clasificación • Modelo Basado en Reglas • Regresión Logística, Arboles Aleatorios y Soporte de Máquina Vectorial
		Diseño de la Evaluación	<ul style="list-style-type: none"> • Estrategia de Validación* • Métricas* 	<ul style="list-style-type: none"> • Validación Cruzada Doble • Exactitud, Precisión, Exhaustividad y Valor F1
		Construcción del Modelo	<ul style="list-style-type: none"> • Configuración de Hiperparámetros 	<ul style="list-style-type: none"> • ver Capítulo 6.3
		Evaluación del Modelo	<ul style="list-style-type: none"> • Resultados del Modelo 	<ul style="list-style-type: none"> • ver Capítulo 6.4
5	Evaluación	Evaluación de Resultados	<ul style="list-style-type: none"> • Mejor Modelo* 	<ul style="list-style-type: none"> • ver Capítulo 6.4

7 Conclusiones y Recomendaciones

7.1. Conclusiones

- En este estudio se desarrollaron tres tipos de modelos para la predicción de hurtos en la ciudad de Medellín basados en aprendizaje de máquinas *Machine Learning*. Los modelos utilizan información histórica del número de delitos en una división espacio temporal e información de la tasa de desempleo. El hurto a personas fue el delito seleccionado, específicamente en la modalidad de atraco, descuido, cosquilleo y raponazo.
- El modelo que presentó las mejores características para la predicción de zonas calientes fue el basado en Máquina de Soporte Vectorial *Support Vector Machine* con un *F1-Score* del 76,06 %, y un Recall del 80 %, por encima de un modelo basado en reglas, una Regresión Logística y uno de Bosques Aleatorios.
- En esta investigación se llevó a cabo un análisis descriptivo y exploratorio de múltiples fuentes de información en la ciudad de Medellín, y se demostró que para el desarrollo de algoritmos de predicción sea confiable, es de gran importancia el adecuado registro de la información por parte de los entes gubernamentales y el procesamiento de los datos, esto con el fin de generar herramientas que hagan posible el despliegue de servicios para ciudades inteligentes.
- Los modelos de predicción de crímenes y de zonas calientes pueden contribuir a un mejor despliegue de los recursos para la seguridad y para tener una mejor respuesta frente a los actos criminales por parte de las autoridades.
- En la presente tesis se desarrolla la metodológica CRISP-DM para la construcción de modelos de predicción de crímenes en la ciudad de Medellín. Dentro de este proceso se definieron estrategias de priorización sobre diversos tipos de delitos, que se podrían replicar en otras ciudades. En este caso, la selección de los parámetros involucrados se realiza con base en la literatura revisada y solo para efectos teóricos.

7.2. Recomendaciones

- Como se pudo evidenciar en los capítulos sobre la preparación de los datos y resultados, la selección de los parámetros para la construcción de los modelos de predicción se

realiza solo para efectos teóricos y se realiza con base en la literatura revisada. Sin embargo, sería adecuado alinear esta selección desde el inicio con las estrategias de las autoridades para la gestión de los recursos de forma más eficiente dentro de la región de estudio. Según la metodología desarrollada, estos parámetros son: el tamaño de la ventana temporal, el tipo de división espacial y su tamaño, los tipos de delitos, y el número de zonas calientes a predecir. Todos estos parámetros se relacionan con los recursos (dinero) disponible que tienen las autoridades, sus estrategias de rotación que tienen implementadas.

- A partir de la construcción de información, se puede observar que la selección de la ventana temporal se definió por la calidad de número de crímenes en cada celda, y debido a la cantidad de registros existentes en Medellín, no fue posible realizar una división menor a una semana. Si se tuviera más información, se podría complementar nuestro conjunto de datos final con características como el día de la semana, el día del mes, incluso a una menor escala, la hora del día. Sin embargo, es claro que existe un gran sub registro de los incidentes por parte de la población, y uno de los principales motivos asociados a esto, es porque las víctimas sienten miedo por represalias que pueda tomar los delincuentes o porque no sienten que las autoridades hagan uso de la información. Es por esto que se señala la necesidad de buscar incentivos y herramientas para que las personas colaboren con los registros de los crímenes y con esto, se puedan mejorar la calidad de las predicciones.
- Con el objetivo de mejorar la predicción de las zonas calientes, se buscará la inclusión de nuevas variables independientes dentro del modelo como zonas especiales en la ciudad de Medellín como: partidos de fútbol, festivales de música o festivales propios de la ciudad, como la Feria de las Flores. La información de las fechas se podría extraer de las páginas web que ofrecen boletería para los eventos de la ciudad y de la alcaldía en donde se anuncian los eventos públicos.
- Para un trabajo futuro se busca completar el ciclo de CRISP-DM, con la implementación de la etapa del despliegue, la cual se encuentra fuertemente ligada al propósito de cada proyecto. Se debe alinear con las herramientas que tienen las autoridades, el presupuesto, la infraestructura y su capacidad de integración. Como se menciona en el capítulo de la metodología CRISP-DM esto puede ser desde un reporte, hasta una tarea repetitiva en tiempo real que se encarga de proveer información para la toma de decisiones.
- Dado que se ha propuesto una metodología general aplicable a la predicción del crimen en cualquier zona, es posible aplicar esta misma metodología en otras ciudades, analizar los resultados obtenidos y compararlos contra los obtenidos en la ciudad de Medellín.

Bibliografía

- [1] Ministerio de Defensa Nacional - Dirección de Estudios Estratégicos, “Información de criminalidad, resultados operacionales y delitos contra las propias tropas.” [En línea] https://www.mindefensa.gov.co/irj/go/km/docs/Mindefensa/Documentos/descargas/estudios_sectoriales/info_estadistica/Avance_Politica_Defensa_Seguridad.xlsx [Último acceso: 2021], Agosto 2021.
- [2] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, *CRISP-DM 1.0*. SPSS Inc., 2000.
- [3] Medellín cómo vamos, “Informe calidad de vida de medellín, 2018.” [En línea] <https://www.medellincomovamos.org/system/files/2020-04/docuprivados/Documento%20ICV%202018.pdf> [Último acceso: 2021], 2019.
- [4] Medellín cómo vamos, “Informe calidad de vida de medellín, 2020.” [En línea] <https://www.medellincomovamos.org/system/files/2021-09/docuprivados/Documento%20Informe%20de%20Calidad%20de%20Vida%20de%20Medell%C3%ADn%202020.pdf> [Último acceso: 2021], 2021.
- [5] D. Yang, T. Heaney, A. Tonon, and L. Wang, “Crimetelescope: crime hotspot prediction based on urban and social media data fusion,” *World Wide Web*, vol. 21(5), pp. 1323–1347, 2017.
- [6] M. W. Yu, C. and Ward, M. Morabito, and W. Ding, “Crime forecasting using data mining techniques,” *2011 IEEE 11th International Conference on Data Mining Workshops*, 2011.
- [7] Y. Zhuang, M. Almedida, M. Morabito, and W. Ding, “Crime hot spot forecasting: A recurrent model with spatial and temporal information,” *2017 IEEE International Conference on Big Knowledge (ICBK)*, pp. 143–150, 2017.
- [8] A. Araujo, N. Cacho, L. Bezerra, C. Vieira, and J. Borges, “Towards a crime hotspot detection framework for patrol planning,” *2018 IEEE 20th International Conference on High Performance Computing and Communications*, 2018.
- [9] A. Rummens, W. Hardyns, and L. Pauwels, “The use of predictive analysis in spatio-temporal crime forecasting: Building and testing a model in an urban context,” *Applied Geography*, vol. 86, pp. 255–261, 2017.

- [10] Y. L. Lin, M. F. Yen, and L. C. Yu, “Grid-based crime prediction using geographical features,” *ISPRS International Journal of Geo-Information*, vol. 7, p. 298, 2018.
- [11] J. Borges, “Time-series features for predictive policing,” *2018 IEEE International Smart Cities Conference (ISC2)*, pp. 1–8, 2018.
- [12] C. Kadar and I. Pletikosa, “Mining large-scale human mobility data for long-term crime prediction,” *EPJ Data Sci*, vol. 7, p. 26, 2018.
- [13] S. K. Dash, I. Safro, and R. S. Srinivasamurthy, “Spatio-temporal prediction of crimes using network analytic approach,” *2018 IEEE International Conference on Big Data*, 2018.
- [14] G. L. Shoesmith, “Space–time autoregressive models and forecasting national, regional and state crime rates,” *International Journal of Forecasting*, volume=29, year=2013, pages=191–201, doi = 10.1016/j.ijforecast.2012.08 .
- [15] O. Kounadi, A. Araujo, and M. Leitner, “A systematic review on spatial crime forecasting,” *Crime Sci*, vol. 9, 2020.
- [16] Arias Sevilla, P., “Pirámide de maslow.” [En línea] <https://economipedia.com/definiciones/piramide-de-maslow.html> [Último acceso: 2021], 25 Febrero 2015.
- [17] OECD, “Better life index, security.” [En línea] <https://www.oecdbetterlifeindex.org/topics/safety/> [Último acceso: 2021].
- [18] PNUD, “Sinopsis: Seguridad ciudadana.” [En línea] <https://www1.undp.org/content/undp/es/home/librarypage/crisis-prevention-and-recovery/IssueBriefCitizenSecurity.html> [Último acceso: 2021], 15 Abril 2014.
- [19] Ministerio de Defensa Nacional, “Marco de convivencia y seguridad ciudadana.” [En línea] https://www.mininterior.gov.co/sites/default/files/politica_marco_de_convivencia_y_seguridad_ciudadana.pdf [Último acceso: 2021], 2019.
- [20] Ministerio de Salud y Protección Social Colombia, “Política nacional de salud mental - resolución 4886 de 2018.” [En línea] <https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/PP/politica-nacional-salud-mental.pdf> [Último acceso: 2021], 2018.
- [21] Congreso de la República, “Ley 599 de 2000 código penal colombiano.” [En línea] http://www.secretariosenado.gov.co/senado/basedoc/ley_0599_2000.html [Último acceso: 2021], 2000.

- [22] S. Quintero, “Citizen security and community participation in latin america,” *Revista Científica General José María Córdoba*, vol. 18, pp. 5–24, 2020.
- [23] J. Laufs and et al., “Security and the smart city: A systematic review,” *Sustainable Cities and Society*, vol. 55, 2020.
- [24] A. Ramaprasad, A. Sanchez-Ortiz, and T. Syn, “A unified definition of a smart city,” *International Conference on Electronic Government*, 2017.
- [25] R. W. Siegfried, “A unified definition of a smart city,” *Cities*, vol. 81, pp. 1–23, 2018.
- [26] M. Bourmpos, A. Argyris, and D. Syvridis, “Smart city surveillance through low-cost fiber sensors in metropolitan optical networks.,” *Fiber and Integrated Optics*,, vol. 33, pp. 205–223, 2014.
- [27] Alcaldía de Medellín, “Sistema de información para la seguridad y convivencia - sisc.” [En línea] <https://www.medellin.gov.co/irj/portal/medellin?NavigationTarget=contenido/8148-Sistema-de-Informacion-para-la-Seguridad-y-Convivencia—SISC> [Último acceso: 2021], 2014.
- [28] Alcaldía de Medellín, “Cámaras de cctv.” [En línea] <https://www.medellin.gov.co/simm/camaras-de-circuito-cerrado> [Último acceso: 2021], 2013.
- [29] Policía Nacional de Colombia, “Audiencia pública de rendición de cuentas 2019.” [En línea] https://www.policia.gov.co/sites/default/files/descargables/informe_audiencia_rendicion_de_cuentas-vig-2019.pdf [Último acceso: 2021], 2020.
- [30] Policía Nacional de Colombia, “Tepillé.” [En línea] <https://tepilleapp.com/> [Último acceso: 2021].
- [31] Real Academia Española, “Delito.” [En línea] <https://dle.rae.es/delito?m=form> [Último acceso: 2021].
- [32] Real Academia Española, “Crimen.” [En línea] <https://dle.rae.es/crimen?m=form> [Último acceso: 2021].
- [33] Real Academia Española, “Hurto.” [En línea] <https://dle.rae.es/hurto?m=form> [Último acceso: 2021].
- [34] Real Academia Española, “Robo.” [En línea] <https://dle.rae.es/robo?m=form> [Último acceso: 2021].
- [35] Policía de Puerto Rico, *Manual de Información Uniforme de Datos del Crimen*. 2006.

- [36] DANE, “Encuesta de convivencia y seguridad ciudadana.” [En línea] <http://microdatos.dane.gov.co/index.php/catalog/574/datafile/F20/V642> [Último acceso: 2021], 2013.
- [37] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer Science+Business Media, 2006.
- [38] A. Samuel, “Some studies in machine learning using the game of checkers,” *IBM Journal of Research and Development*, vol. 3, pp. 210–229, 1959.
- [39] A. Geron, *Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow - Concepts, Tools, and Techniques to Build Intelligent Systems*. Sebastopol: O'Reilly, 2019.
- [40] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2 ed., 2003.
- [41] C. J. C. Burges, “A tutorial on support vector machines for pattern recognition,” *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.
- [42] FAYRIX, “Selección de métricas para aprendizaje automático.” [En línea] https://fayrix.com/machine-learning-metrics_es [Último acceso: 2021].
- [43] P. Gupta and N. K. Sehgal, *Introduction to Machine Learning in the Cloud with Python: Concepts and Practice*. Cham, Switzerland: Springer International Publishing, 2021.
- [44] Z. Somogyi, *The Application of Artificial Intelligence: Step-by-Step Guide from Beginner to Expert*. Cham, Switzerland: Springer, 2021.
- [45] J. Browniee, “Train-test split for evaluating machine learning algorithms.” [En línea] <https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/> [Último acceso: 2021], 24 Julio 2020.
- [46] J. Browniee, *Machine Learning Algorithms From Scratch: with Python*. Machine Learning Mastery, 2016.
- [47] J. Browniee, “A gentle introduction to k-fold cross-validation.” [En línea] <https://machinelearningmastery.com/k-fold-cross-validation/> [Último acceso: 2021], 23 Mayo 2018.
- [48] J. Browniee, “Nested cross-validation for machine learning with python.” [En línea] <https://machinelearningmastery.com/nested-cross-validation-for-machine-learning-with-python/> [Último acceso: 2021], 29 Julio 2020.
- [49] S. Raschka and V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2, 3rd Edition*. Packt Publishing, 2019.

- [50] T. Lin, Y. Xie, A. Wasilewska, and C. Liau, *Data Mining: Foundations and Practice*. Studies in Computational Intelligence, Springer Berlin Heidelberg, 2008.
- [51] D. Larose, *Data Mining and Predictive Analytics*. Wiley Series on Methods and Applications in Data Mining, Wiley, 2015.
- [52] D. Olson and D. Delen, *Advanced Data Mining Techniques*. Springer Berlin Heidelberg, 2008.
- [53] P. J. Brantingham and B. P. L, *Patterns in Crime*. New York: Macmillan, 1984.
- [54] J. Tiihinen, P. Halonen, L. Tiihonen, K. H., M. Storvik, and J. Callaway, “The association of ambient temperature and violent crime,” *Sci Rep*, 2017.
- [55] P. Butke and S. Sherida, “An analysis of the relationship between weather and aggressive crime in cleveland, ohio,” *Weather, Climate and Society*, vol. 2, pp. 127–139, 2010.
- [56] L. Alves, H. Ribeiro, and F. Rodrigues, “Crime prediction through urban metrics and statistical learning,” *Physica A: Statistical Mechanisc and its Applications*, vol. 505, pp. 435–443, 2018.
- [57] M. Andresen and N. Malleson, “Intra-week spatial-temporal patterns of crime,” *Crime Sci*, vol. 3, 2015.
- [58] M. Williams, P. Burnap, and L. Sloan, “Crime sensing with big data: The affordances and limitations of using open-source communications to estimate crime patters,” *The British Journal of Criminology*, vol. 57, pp. 320–340, 2017.
- [59] A. A. Biswas and S. Basak, “Forecasting the trends and patterns of crime in bangladesh using machine learning model,” in *2nd International Conference on Intelligent Communication and Computational Techniques (ICCT)*, 2019.
- [60] A. Bogomolov, B. Lepri, J. Staiano, F. Oliver, N. Pianesi, and A. Pentland, “Once upon a crime: Towards crime prediction from demographics and mobile data,” in *Poceedings of the 16th International Conference on Multimodal Interaction ICMI*.
- [61] A. A. Junior, N. Cacho, A. C. Thome, A. Medeiros, and J. Borges, “A predictive policing application to support patrol planning in smart cities,” in *International Smart Cities Conference (ISC2)*, 2017.
- [62] Secretaria de Seguridad de Bogotá, “Secretaría seguridad bogotá - diseño y validación de modelos de analítica predictiva de fenómenos de seguridad y convivencia para la toma de decisiones en bogotá.” [En línea] <https://www.facebook.com/secretariadesseguridadbogota/videos/2683121515289433/?t=8>. [Último acceso: 2021], 2020.

- [63] Alcaldía de Medellín, “Epm desarrolla el sistema de información para análisis de entorno consumiendo datos públicos del portal medata.” [En línea] <http://medata.gov.co/historia/epm-desarrolla-el-sistema-de-información-para-análisis-de-entorno-consumiendo-datos>. [Último acceso: 2021], 2018.
- [64] V. Munoz, M. Vallejo, and J. E. Aedo, “Machine learning models for predicting crime hotspots in medellin city,” in *2021 2nd Sustainable Cities Latin America Conference (SCLA)*, 2021.
- [65] Secretaría de Seguridad y convivencia - Sistema de Información para la Seguridad y la Convivencia SISC, “Hurto a persona.” [En línea] <http://medata.gov.co/dataset/hurto-persona> [Último acceso: 2021], 2020.
- [66] Pydata, “pandas.” [En línea] <https://pandas.pydata.org/> [Último acceso: 2021].
- [67] Alcaldía de Medellin, “Límite catastral de comunas y corregimientos.” [En línea] https://geomedellin-m-medellin.opendata.arcgis.com/datasets/283d1d14584641c9971edbd2f695e502_6 [Último acceso: 2021].
- [68] Departamento Administrativo Nacional de Estadísticas, “Información histórica del mercado laboral.” [En línea] <https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo-y-desempleo/mercado-laboral-historicos> [Último acceso: 2021], 2021.
- [69] SIATA, “Generalidades de la información red meteorológica y pluviométrica del valle de aburrá.” [En línea] https://siata.gov.co/descarga_siata/index.php/info/pluviomet/ [Último acceso: 2021].
- [70] Área Metropolitana del Valle de Aburrá, “Preparémonos para el inicio de la primera temporada de lluvias de 2019.” [En línea] <https://www.metropol.gov.co/Paginas/Noticias/preparemonos-para-el-inicio-de-la-primer-temporada-de-lluvias-de-2019.aspx> [Último acceso: 2021], 2019.
- [71] D. Wang, W. Ding, H. Lo, T. Stepinski, J. Salazar, and M. Morabito, “Crime hotspot mapping using the crime related factors—a spatial data mining approach,” *Applied Intelligence*, vol. 39(4), p. 772–781, 2012.
- [72] V. Munoz, M. Vallejo, and J. E. Aedo, “Exploratory analysis of crime behavior in the city of medellin,” in *2021 2nd Sustainable Cities Latin America Conference (SCLA)*, 2021.
- [73] J. Browniee, “A gentle introduction to imbalanced classification.” [En línea] <https://machinelearningmastery.com/what-is-imbalanced-classification/> [Último acceso: 2021], 23 Diciembre 2019.

-
- [74] Scikit-Learn, “Logistic regression classifier.” [En línea] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html [Último acceso: 2021].
 - [75] Scikit-Learn, “Random forest classifie.” [En línea] <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html> [Último acceso: 2021].
 - [76] Scikit-Learn, “C-support vector classification.” [En línea] <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html> [Último acceso: 2021].